時間周波数表現による音信号混合法

大脇 渉

電気通信大学
情報理工学研究科

博士 (工学) の学位申請論文

2016 年 3 月

時間周波数表現による音信号混合法

博士論文審査委員会

主査　高橋 弘太　准教授
委員　張 熙　　教授
委員　肖 鳳超　　教授
委員　野村 英之　准教授
委員　三橋 渉　　名誉教授

The sound signal mixing method using the time-frequency plane

Wataru OWAKI

Abstract

This study has two purposes. The first purpose is the realization of a sound signal mixing method, which can mix a voice and background music without hiding the voice. The second purpose is to show the effectiveness of using the time-frequency plane for the sound signal mixing. In this paper, a sound signal mixing method using the time-frequency plane is proposed. The main feature of the proposed method is that the proposed method has focused on two patterns on the time-frequency plane of the voice i.e. the harmonic structure and the formant structure. By utilizing these two structures, the proposed method can be used in every mixing ratio for every type of voice. In order to evaluate of the proposed method, a comparison between the proposed method, the equalizer approximation of the proposed method, and the ducker approximation of the proposed method, was carried out by the subjective evaluation experiments and observation of the time-frequency planes of outputs of these processes. The results of these evaluations had shown that the proposed method can achieve the purposes of this study.

概 要

本研究の目的は, 第一に, 音声と BGM(Background Music) を入力した時に音声を埋もれさせない音信号混合を実現することであり, 第二に, 音信号混合に時間周波数平面を用いることの有効性を示すことである. ここで, 音声を埋もれさせない音信号混合を, 第一に音声の内容の聴き取りを維持するものとし, 第二に音声と BGM の音質を維持するものとする. 本論文では, これらの目的を達成するために, 時間周波数平面を用いた音信号混合法を 3 手法提案し, 主観評価実験と時間周波数平面の観察により有効性を検討した.

第 2 章で, 音楽制作におけるミキシング処理についてまとめた. まず, ミキシング処理の目的が, 音量, 定位, 音色のバランスを整え, 主役をはっきりさせることにあるとまとめた. 続いて, 従来のミキシング処理として 5 例を例示し, 従来のミキシング処理が時間領域処理と周波数領域処理の組み合わせによる時間周波数領域処理にとどまっていることを示した. その問題点として, 時間領域処理と周波数領域処理の組み合わせでは実現できない時間周波数領域処理がある点を指摘した.

第 3 章で, スマートミキサーについて述べた. スマートミキサーの定義は, 時間周波数平面を用いた解析部をミキサーの内部に持ち, ミキシングを時間周波数平面上での重ね合わせとして実施する音信号のミキサーである. スマートミキサーとしての利点は, 従来のミキシング処理では操作者に頼っていた聴感的な条件の判定をミキサー自体が担うことができ, さらに処理内容に反映できる点にある. また, スマートミキサーの実装について, ハードウェアとソフトウェアについて例示した.

第 4 章で, 調波構造に着目した手法を提案法 A として提案した. この手法では, 時間周波数平面上での調波構造のパターンを捉えるために, MPEG1 Audio に含まれる聴覚心理モデルの純音判定を基準とした. この提案法 A について, 出力信号の時間周波数平面の観察と聴取実験による主観評価実験を行った. 評価実験では, 提案法 A の時間周波数平面上での処理量を時不変のイコライザとして近似する処理法を比較対象とした. 比較により, 音声の聴き取りやすさについて, 提案法 A の評価が上回る (ウィルコクソンの順位和検定の片側検定において, $p = 0.027$) ことがわかった.

第 5 章で, フォルマント構造に着目した手法を提案法 B として提案した. この手法では, 音声の母音の識別性を担うフォルマント構造を基準とした. フォルマント構造は調波構造と比較して時間周波数平面上での領域が狭いことから, 聴きとりと音質を両立できると考えた. 主観評価実験によって提案法 B の有効性を確かめるとともに, 子音のうち特に歯擦音に聴き取りづらさが残ると考察した.

第 6 章で, フォルマント構造と歯擦音に着目した手法を提案法 C として提案した. この手法は, 提案法 B によるフォルマント構造への処理に, 歯擦音への処理を追加したものである. 歯擦音への処理は, 母音の周波数分布の特徴は 3 kHz より低い周波数帯域で顕著であり, 歯擦音の周波数分布の特徴は 4 kHz よりも高い周波数帯域で顕著であることから, 4 kHz より高い周波数帯域において歯擦音のスペクトル遷移パターンを強調するものとした. 同一女性話者による子音のみが異なる 3 種の音声信号を用いた実験により, 提案法 C が適切に時間周波数平面上での歯擦音のパターンを強調できることを示した.

第 7 章では, 提案した 3 手法について比較と考察を行った. 主観評価実験により, 音声の聴きとりやすさと音質について提案法 C が最良であることがわかった. また, 提案法 A の計算量が最小であり, 提案法 C と同等の音声の聴きとりやすさの評価が得られることがわかった. 一方, 音声符号化形式を用いる場合では, 提案法 B が有用である.

第 8 章では, 音信号混合法への時間周波数平面利用の有効性を確認した. 提案法 A を周波数領域処理および時間領域処理として近似した手法を定義し, 時間周波数平面での観察と客観的な数値指標で比較した. 比較によって, 音声の聴きとりやすさと音質の両方が提案法 A によってのみ達成できることを明らかにし, 時間周波数領域での処理が有効であることを示した.

以上より, 本論文で提案する手法により, 音声が埋もれない音信号混合を実現できることが示された. また, 音信号混合に時間周波数平面を用いることの有効性も示された.

目 次

表 1. 本論文での記号の定義

F_{s}	サンプリング周波数
n	サンプル番号
$x_{\mathrm{A}}[n]$	優先度が高い入力信号の振幅値
$x_{\mathrm{B}}[n]$	優先度が低い入力信号の振幅値
$y[n]$	出力信号の振幅値
i	時間フレーム番号
k	周波数ビン番号
$X_{\mathrm{A}}[i,k]$	$x_{\mathrm{A}}[n]$ の時間周波数平面上の位置 (i,k) での, 短時間フーリエ変換値
$X_{\mathrm{B}}[i,k]$	$x_{\mathrm{B}}[n]$ の時間周波数平面上の位置 (i,k) での, 短時間フーリエ変換値
$Y[i,k]$	$y[n]$ の時間周波数平面上の位置 (i,k) での, 短時間フーリエ変換値
$L_{\mathrm{A}}[i,k]$	$X_{\mathrm{A}}[i,k]$ の聴感上での音量
$L_{\mathrm{B}}[i,k]$	$X_{\mathrm{B}}[i,k]$ の聴感上での音量
$W_{\mathrm{A}}[i,k]$	$X_{\mathrm{A}}[i,k]$ へのゲイン係数
$W_{\mathrm{B}}[i,k]$	$X_{\mathrm{B}}[i,k]$ へのゲイン係数
$w_{\max}$	最大抑制量
δ	聴覚上の音量 $L_{\mathrm{A}}[i,k]$,$L_{\mathrm{B}}[i,k]$ の音量差
$M_{\mathrm{B}}[i,k]$	$X_{\mathrm{B}}[i,k]$ への同期係数
$\phi_{\mathrm{A}}[i,k]$	$X_{\mathrm{A}}[i,k]$ の位相
$\phi_{\mathrm{B}}[i,k]$	$X_{\mathrm{B}}[i,k]$ の位相
$\mathrm{ATH}[k]$	最小可聴値
$S_{\mathrm{A}}[i,k]$	$X_{\mathrm{A}}[i,k]$ の聴覚上の音量
$S_{\mathrm{B}}[i,k]$	$X_{\mathrm{B}}[i,k]$ の聴覚上の音量
$Y_{\mathrm{A}}[i,k]$	ゲイン係数 $W_{\mathrm{A}}[i,k]$ 適用後の $X_{\mathrm{A}}[i,k]$
$Y_{\mathrm{B}}[i,k]$	ゲイン係数 $W_{\mathrm{B}}[i,k]$ 適用後の $X_{\mathrm{B}}[i,k]$
$L_{\mathrm{dB}}[X,i,k]$	$X[i,k]$ の dB 値
$C_{\mathrm{tm}}[X,i,k]$	$X[i,k]$ への純音判定
$C_{\mathrm{ATH}}[X,i,k]$	$X[i,k]$ への有音判定
$L^{A}[X,i]$	$X[i,k]$ の全周波数ビンでの平均エネルギ
$L^{A}_{\mathrm{dB}}[X,i]$	$L^{A}[X,i]$ の dB 値
D_{tm}	有音な純音成分
D_{tn}	有音な純音成分の近傍領域成分
D_{nm}	有音な非純音成分
D_{na}	非可聴でかつ近傍領域内に有音な純音が存在しない領域
$F_{\mathrm{lim}}(g,l,u)$	g を上限 u と下限 l の間の値に制限するリミット関数
$L^{f}[X,i,k]$	時間周波数表現の近傍周波数帯での平均化パワー

$E_{\mathrm{A}}[i,k]$	$X_{\mathrm{A}}[i,k]$ のスペクトル包絡
$D_{\mathrm{G}}[i,k]$	ゲイン操作実行判定
$P_{\mathrm{A}}[i,k]$	$X_{\mathrm{A}}[i,k]$ のパワー
$P_{\mathrm{B}}[i,k]$	$X_{\mathrm{B}}[i,k]$ のパワー
$D_{\mathrm{fn}}[i,k]$	フォルマント帯域判定
$R_p[i]$	時間フレーム毎の入力信号のパワー比率
α	$W_{\mathrm{B}}[i,k]$ のフレーム間での減衰係数
L_α	$G_{\mathrm{B}}[i,k]$ の下限値
N_{LPC}	LPC 次数
w_{fmt}	フォルマント帯域幅
$D_{\mathrm{G}}^{\mathrm{S}}[i,k]$	歯擦音に特化したゲイン係数のゲイン操作実行判定
$W_{\mathrm{A}}^{\mathrm{S}}[i,k]$	優先音 $X_{\mathrm{A}}[i,k]$ への歯擦音に特化したゲイン係数
$W_{\mathrm{B}}^{\mathrm{S}}[i,k]$	非優先音 $X_{\mathrm{B}}[i,k]$ への歯擦音に特化したゲイン係数
$W_{\mathrm{A}}^{\mathrm{F}}[i,k]$	優先音 $X_{\mathrm{A}}[i,k]$ への母音に特化したゲイン係数
$W_{\mathrm{B}}^{\mathrm{F}}[i,k]$	非優先音 $X_{\mathrm{B}}[i,k]$ への母音に特化したゲイン係数
$y_{\mathrm{A}}[n]$	$Y_{\mathrm{A}}[i,k]$ の逆短時間フーリエ変換値
$y_{\mathrm{B}}[n]$	$Y_{\mathrm{B}}[i,k]$ の逆短時間フーリエ変換値
$V_{\mathrm{A}}[i,k]$	$y_{\mathrm{A}}[n]$ の短時間フーリエ変換値
$V_{\mathrm{B}}[i,k]$	$y_{\mathrm{B}}[n]$ の短時間フーリエ変換値
$G_{\mathrm{c}}^{\mathrm{E}}[k]$	イコライザ近似でのゲイン係数
$Z^{\mathrm{E}}[i,k]$	イコライザ近似での出力信号の時間周波数表現
$G_{\mathrm{c}}^{\mathrm{D}}[i]$	ダッカー近似でのゲイン係数
$Z^{\mathrm{D}}[i,k]$	ダッカー近似での出力信号の時間周波数表現

第 1 章
序論

1.1 研究の背景

本論文では, スマートミキサーの構成法を提案する. スマートミキサーの定義は, ミキサー内部に時間周波数平面による解析部を持ち, ミキシングを時間周波数平面の重ね合わせとして実施する音信号のミキサーである. スマートミキサーの利点は, 時間周波数平面の導入によって, 従来のミキシング処理では操作者に頼っていた聴感的な条件の判定をミキサー自体が担うことができ, さらに処理内容に反映できる点にある.

スマートミキサーを研究する背景として, まず音信号のミキシングの研究の重要性が高まりつつある状況について述べる.

近年, 音を活用した機器が増えつつある. 音を活用した機器として, カーナビゲーションシステム, 携帯音楽再生機, スマートフォン, パーソナルコンピューター, テレビ, カラオケ, ワイヤレススピーカー [1] などの機器が該当する. これらの機器と機器で再生される音を, 表 1.1 にまとめる. 各機器には, 複数の音信号が入力され得る. 各機器に複数の音信号が同時に入力された時, これらの機器に内蔵されているミキサーができる処理は, 出力する入力信号の切り替えや音量調整などの簡易な処理である.

次に, 高度なミキシング処理が必要な理由について述べる.

音信号のミキシングとは, 任意数の入力音信号を任意数の出力音信号にまとめる処理である. ここで, 二入力を一出力へと混合する場合を考える. 単純な処理は, 入力信号の各サンプル値の加算を, 同時刻の出力信号のサンプル値とする. しかし, この処理によって得られた出力信号を聴いた時, 各入力信号を単体で聴いた時に得られる聴感の全てを得られるとは限らない. その理由として, 第一に入力信号間の干渉, 第二に人間の聴覚特性が考えられる.

表 1.1. 音を活用した機器と, その機器で主に再生される音

種類	再生する音
カーナビゲーションシステム	カーステレオ, ナビ音声
携帯音楽再生機	音楽, ゲーム, アラーム
スマートフォン	通話音, メール着信音, ゲーム, 音楽
パーソナルコンピューター	動画, IP 電話, メール着信音, 各種アラート
テレビ	テレビ番組, 番組メタ情報
カラオケ	伴奏, 歌声
ワイヤレススピーカ [1]	音楽, プッシュトーク

まず, 入力信号間の干渉について述べる. 各入力信号は, 短時間に局在する単一周波数の正弦波信号の素片の集合体として捉えることができる. 正弦波は, 周波数, 振幅, 位相のパラメータからなる. 複数の正弦波信号を加算したとき, 振幅, 位相は様々に変化する. このため, 各入力信号単体で知覚できた音が知覚できなくなる場合がある.

次に, 人間の聴覚特性について述べる. 聴覚特性として, 時間周波数平面での近接領域に滲んで知覚されることが知られている. 時間周波数平面での滲みによって, 近接領域に分布する複数成分の弁別ができなくなる. 特に, 近接領域内でエネルギの大きな成分が, エネルギの小さな成分を弁別できなくさせる現象をマスキングという. ミキシングは, 1 つの音信号に複数の音信号の成分を詰め込むことに相当する. 詰め込むことで, 単体では判別することができた成分が他の成分に埋もれ, 判別が難しくなる.

これらの理由により, 入力された音信号の成分のうち知覚できる量をできるだけ落とさずに出力信号にまとめるためには, 人間の聴覚特性を考慮した高度なミキシング処理が必要である.

現在, 高度なミキシングを実施している典型例は, 音楽, 映画などの商業的なコンテンツ制作や, テレビ, ラジオなどの放送局である. これらの現場では, 堅牢な防音処理が施されたスタジオに, 巨大なミキシングコンソールをはじめとした多種多様な音響装置を保有する. さらに, これらの音響装置は, 一度構築すれば完成ではない. ミキシングする音が変化したときに組み換えが必要となる. これらの装置の組み換えを行い, それを操作するには, 音響装置の特性とミキシング対象との相性を熟知したプロフェッショナルが必要である.

このように, 高度なミキシングを実現するためには, 機材でも人材の面でも莫大なコストがかかる. 高度なミキシングを生活に取り入れていくためには, 技術革新が必要である.

一方, 一般人にとって音を用いた作業は身近になりつつある.

第一の理由として, よい機材を個人が入手しやすくなったことがある. パーソナルコンピューターの高性能化と低価格化によって, 従来のミキシング法に迫る処理を, 個人

所有の PC で実現可能になりつつある. さらに近年, 携帯音楽再生機やスマートフォン用のアプリケーションとして, 音楽制作のミキシングアプリケーションが登場しはじめた.

第二の理由として, コンテンツ制作のモチベーションの高まりがある. 企業や学校でのプレゼンテーションでは, デモンストレーションの動画制作が一般化し, 音の編集がついてまわる. さらにはプライベートでも, 結婚式や誕生日などの機会に, ビデオレターを制作することが盛んになりつつある.

よい機材が入手できると, 今度はミキシングのノウハウが必要になる. しかし, 技術の蓄積を行い, 生活の様々な場面で調整に労力をかけるのは現実的ではない. 高度なミキシング処理自体の敷居を下げ, 効率的に実施できる技術が必要である.

そこで, スマートミキサーの登場である.

スマートミキサーが従来のミキサーと比較してスマートである点は, 入力信号を解析する機構をミキサー自体が持っている点にある. 解析する機構にプロフェッショナルのノウハウと音響信号処理の技術を結集したアルゴリズムを搭載することで, 従来の高度なミキシングを手軽に実現できる.

スマートミキサーによって, 従来のミキシングで行われてきた処理の緻密さをさらに押し広げることができる. スマートミキサーでは, 解析する機構として入力信号を時間周波数平面に展開する. 一方, 従来のミキシングで用いられている機器は, 時間領域での処理が大半である. 時間周波数平面で信号を捉えることで, 従来のミキシングでは実施が難しかった, 緻密な周波数分解能での処理が可能である.

このように, スマートミキサーは一般からプロフェッショナルまで, リビングからスタジアムまで, 幅広く音世界を豊かにできる可能性を持った研究である. スマートミキサーによる社会貢献について, 3.3.3 節で詳述する.

本論文は, 音声と BGM をミキシングした際に, 音声を埋もれさせない音信号混合を実現するスマートミキサーの構成について検討したものである. 音声を埋もれさせない音信号混合とは, 音声の内容の聴き取りを第一の目的とし, 音声と BGM の音質の維持を第二の目的とする. 提案法では, 音声を埋もれさせない音信号混合を実現するための方策として, 音声の時間周波数平面上での特徴である, 聴覚心理モデルと音声生成モデルの導入を検討している.

本論文の成果により, 音を活用する機器にスマートミキサーを実装する試みに弾みがつく.

1.2　従来研究

本節では, ミキシングの従来研究について紹介する.

ミキシングの従来研究は, 大きく 3 つのアプローチに分類できる. 第一に, パラメータの自動調整である. 第二に, 聴感特性に着目したテレビ音声のミキシングである. 第三に, 仮想空間のモデリングにおける聴感特性に着目したレンダリングである.

いずれのアプローチも, 従来のミキシング構成は保持しつつパラメータ調整を自動化する試みである. 従って, これらのアプローチはプロフェッショナルの緻密な処理を実現する難しさを解消するものではない. 以下, 3 つのアプローチそれぞれについて述べる.

1.2.1　パラメータの自動調整

代表的な研究として, 関西学院大学の谷井ら [2] によるミックスダウンデザインテンプレート, イギリスのクイーンメリー大学の Perez - Gonzalez ら [3] による AM-DAFx が挙げられる. ここで, 自動調整の対象とする一般的なパラメータとは, 音量, 左右の音量差, パラメトリックイコライザの中心周波数, Q 値である.

ミックスダウンデザインテンプレート [2] は, プロが制作したミキシングの設定を, 楽曲のジャンル別にテンプレート化する手法である. ユーザは, 制作する楽曲の目指すジャンルに合わせて, テンプレートを選択し適用する.

この手法の主眼は, ノウハウの蓄積と再利用にある. まず, ミックスダウンデザインとは, ミキシングエンジニアがミキシング作業時に設定する, エフェクト処理の順序とパラメータ, 各トラックデータのボリュームバランス, パンニングなどの設定情報である. また, 楽曲の A メロ, B メロ, サビなどの構成を, 楽曲構成グループと呼ぶ. このとき, ミックスダウンデザインを楽曲構造グループごとにテンプレート化し, 再利用可能な形にしたものがミックスダウンデザインテンプレートである. このテンプレートを任意の楽曲構造グループに適用することで, 比較的容易に最適なミキシングが得られるとしている. テンプレート化することで, プロフェッショナルのノウハウを蓄積し, 再利用することができる. このため, 制作時間の短縮につながり, 作品の生産性に寄与できるとしている.

AM-DAFX(automatic-mixing degital effect tools)[3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15] は, パラメータを適応的アルゴリズムで自動調整するものである. 通常のリアルタイムでのミキシングでは, ミキシングエンジニアが処理結果をモニタリングすることで, 未来の処理結果を予測し, パラメータを調整していく. このミキシングエンジニアの役割を, 適応的アルゴリズムで置き換える方針である.

1.2.2　聴感特性に着目したテレビ音声のミキシング

テレビ音声でのミキシングに関する従来研究として, 音声と背景 BGM の音量のバランスに着目したものが挙げられる. 高齢者になると聴感特性が変化する. この聴感特性に着目した高齢者向けのミキシング装置の開発が研究の中心である.

代表的な研究として, 早稲田大学の山崎ら [16, 17] による研究と, NHK 技研の小森ら [18] による研究が挙げられる.

栗田ら [16] は, 近年の高齢化や生活様式の多様化に着目し, 多様な受聴環境を考慮したテレビ音声のミキシングについて研究, 調査を行っている. この研究では, 被験者が自

らナレーションと BGM との音量バランスを操作できることに好感を持つとしている.

山崎ら [17] は, 高齢者に聴こえやすい放送音声サービスの研究の中で, 背景音の音響的特徴と聞きやすさの検討と, 放送番組音声の背景音レベルと聞きやすさの検討を行っている. 高齢者と非高齢者での音量バランスの好みの違いについて実験から, 視聴者の音量バランスの好みに合わせて試聴できる配信方法を検討すべきだとしている.

小森ら [18] は, ナレーションと背景音のミキシングバランスに関して, ラウドネスレベルの相対値を用いて検討した. 算出結果を利用し, ナレーションと背景音のミキシングバランスが適正であるかどうかを表示する装置を試作している. 番組のジャンルごとに最適なバランスが異なるとし, 情報系の番組での適正なラウドネスレベルの相対値を決定している.

ここで, テレビ放送でのミキシングと本研究が主眼とするミキシングでは, 発想に違いがある. テレビ放送では, 音声の聴き取りが最重要であり, BGM の音量を含めた音質は重視されていない. 一方, 本研究が主眼とする音声を埋もれさせない音信号混合では, 音声と BGM の音質を両立させるものとしている.

1.2.3　仮想空間のモデリングにおける聴感特性に着目したレンダリング

ここでの仮想空間とは, 例えば戦闘ゲームの戦場である. 戦闘ゲームでは, 戦闘意欲を高める豪奢な BGM, 勇ましい声優の音声に加えて, 四方八方に飛び交う銃弾の飛行音, 銃声, 怒声, 悲鳴などの多種多様な効果音をミキシングする必要がある. しかも, 遅延を最小限に抑えられなければ, ゲームの操作感が損なわれてしまう. 効率良く数多くの音信号をリアルタイムで処理する必要がある.

代表的な研究として, フランスの INRIA の Tsingos ら [19, 20, 21, 22] による scalable perceptual mixing, ギリシャの University of Patras の Tsilfidis ら [23] による hierarchical perceptual mixing, ポーランドの AGH University of Science and Technology の Kleczkowski ら [24, 25, 26, 27, 28, 29] による selective mixing が挙げられる.

これらの研究では, 聴感特性の一つであるマスキングに着目することで, ゲームにおける音処理への要求に応えている. 入力信号の成分のうち, より大きく聴こえる成分のみを階層的に選択していく. 階層化によって効率良く再生する信号を選別し, 出力信号を組み上げる.

1.3　研究目標

本研究の目標は, プロフェッショナルをも超える緻密な処理を, 手軽に実現することである. ミキシングのプロフェッショナルに, これまでにない緻密さと自由な処理を提供する.

本論文では, 検討するミキシングの対象を, 音声と BGM が入力信号のときであるとし, 音声を埋もれさせない音信号混合を実現する手法を検討する. 音声を埋もれさせない音信号混合を, 音声の内容の聴き取りを第一の目的とし, 音声と BGM の音質の維持

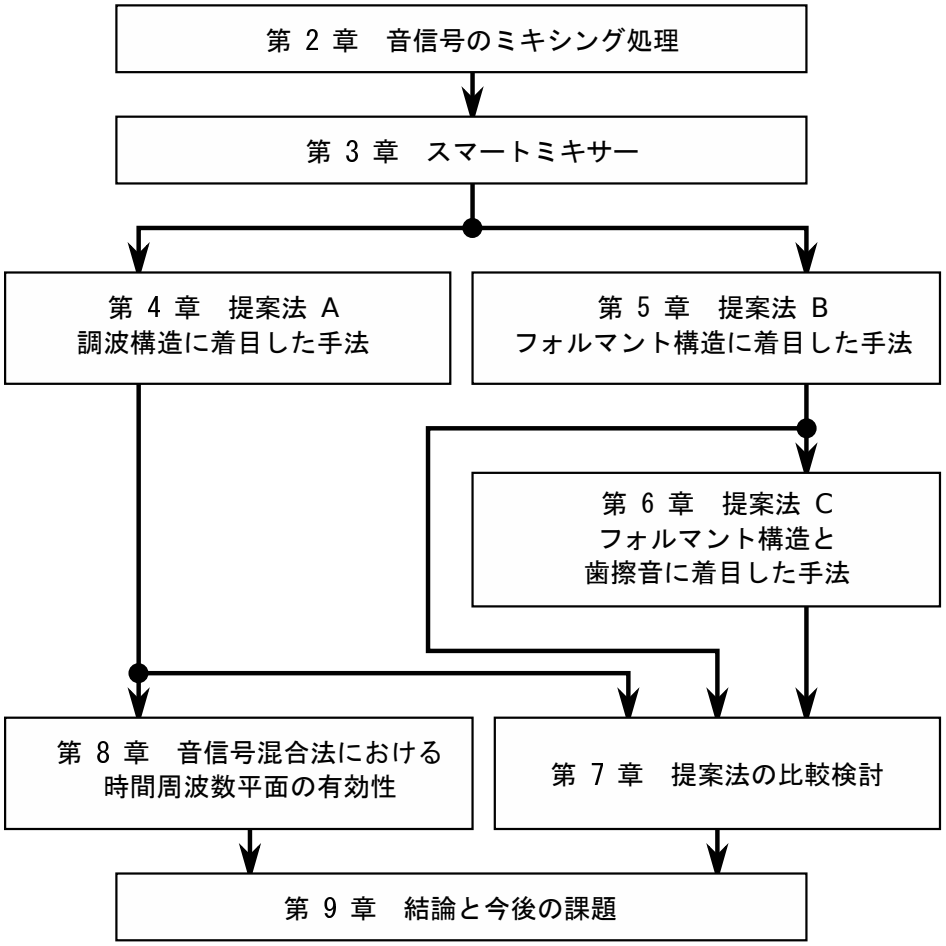


図 1.1. 本論文の構成

を第二の目的とする. ここで, 音声や BGM の音質の壊れよりも, 音声の内容の聴き取りを優先する. 実用のためには, リアルタイムで処理が行うことできるアルゴリズムである必要がある. できるだけ算出手間が少なく, 参照する過去と未来のサンプル数が少ない処理を目指す.

加えて, パラメータ調節を行わなくとも, 入力時の音量や声質に違いがあっても素材を選ばず, 満足のいく程度の処理ができるアルゴリズムの構築を目指す. また, 用意するパラメータは, 直感的な調整ができるようなものとすることを目指す.

さらに, 完成したアルゴリズムは, 音楽制作を始めとするコンテンツ制作の従来機材への導入が行える方式での実装や, 様々な機器に搭載可能なリアルタイム動作する機器として実装したい.

1.4　本論文の構成

本論文の構成を, 図 1.1 に整理して示す.

まず, 第 2 章で音楽制作でのミキシング処理について述べる.

次に, 第 3 章でスマートミキサーについて述べる.

第 4 章で調波構造に着目した手法を提案法 A として提案する. まず, 第 4.1 節で音量関係に着目した手法での検討について述べる. 時間周波数平面上での処理として, 振幅処理と位相処理が考えられる. 初めにそれぞれの処理について複数のパラメータによる単純な処理の構成法を検討し, 所望の効果が得られる処理条件を考察した. 続いて振幅処理と位相処理との連携法を検討した. 検討で得られた知見として, 入力信号の調波構造のスペクトル遷移パターンを強調するのが効果的であった. この知見を元に, 第 4.2 節で聴覚心理モデルに基づいた手法を検討する. 調波構造のスペクトル遷移パターンを少ないパラメータで捉える目的で, 聴覚心理モデルの導入を検討した. まず, 振幅処理と位相処理についてそれぞれ検討し, 続いて連携法を検討した. 検討の結果, 音声の聴き取りは確保できている一方で, 音質劣化が過度であるとの評価であった. 第 4.3 節で提案法 A について述べる. 提案法 A は時間周波数平面上でなだらかなゲインマスクを生成する音質に配慮した手法である. 聴取実験により, 3 組の音信号について同一パラメータを用いて音声の聞き取りを確保できることを確認した.

第 5 章で, フォルマント構造に着目した手法を提案法 B として提案する. フォルマント構造は音声の母音識別に用いられる時間周波数平面上の特徴量である. 検討の結果, 母音の識別性を確保できている一方で, 子音の中でも歯擦音の聴き取りやすさが不十分であることがわかった.

第 6 章で, フォルマント構造と歯擦音に着目した手法を提案法 C として提案する. この手法は, 提案法 B によるフォルマント構造への処理に, 歯擦音への処理を追加したものである.

第 7 章で, 3 つの提案法を比較検討する.

第 8 章で, 音信号混合法における時間周波数平面利用の有効性を, 提案法 A について検討する.

第 9 章で結論を述べる.

第 2 章
音信号のミキシング処理

この章では, 音信号のミキシング処理の概略について考察する. 特に, 音楽製作でのミキシングを中心に紹介する. また, ミキシング処理時に考慮する聴感特性についても紹介する.

2.1 ミキシングとは

ミキシングとは, ダウンミキシングの略称として用いられることが一般的である. ダウンミキシングとは, 複数チャンネルの入力信号を, 任意のチャンネル数にまとめて落としこむ作業 [30] のことである.

ミキシング (mixing) と同義の用語には, ミクシング (mixing), ミックス (mix), ミックスダウン (mix down), トラックダウン (truck down, TD)[31] が挙げられる. フィルム制作では, リレコーディング, ダビングと呼ぶ. テレビ制作では, スイートニングとも呼ぶ [32].

本稿では, ミキシングに統一して表記する.

2.2 ミキシングの役割

この節では, 視聴覚コンテンツの製作過程の中における, ミキシングの役割について考察する.

まず, ミキシングの役割について, プロフェッショナルであるミキシングエンジニアの記述を列挙する.

葛巻 [33] は, まず音楽製作の作業についてレコーディング, ミキシング, マスタリングの 3 つに大きく分かれると述べている. その中でミキシングが, 楽曲をまとめ上げる作業として非常に重要であると述べている.

Izhaki[34] は, マルチトラックな素材をバランスのとれた状態にし, マルチチャンネルのある形式にまとめる処理であると述べている. 素材をまとめる上で, どのくらいその音が重要かが有益な問いかけであると述べている.

山内 [35] は, 音楽でのミキシングの目的として, 2 つ挙げている. 第一に, 楽曲にとって主役は何かをはっきりさせること. 第二に, マスキングを回避していくことである. この 2 つの目的からミキシングの方針として, 主役がはっきりしていれば, それ以外のパートに遠慮してもらうことで主役の魅力を損なわずにミックスできると述べている. ここでマスキングとは, ある音が別の音を聞こえなくする, あるいは聞こえにくくする現象 [36] で, 聴覚特性の一種である.

以上の記述から, 本稿ではミキシングの役割を, 次の 3 点とする. 第一に, 任意のチャンネル数に落とし込む. 第二に, 全体を聴こえやすく調整する. 第三に, 主役をはっきりさせる. この 3 点に着目し, ミキシング処理について考察する.

2.3　ミキシングの目的

ミキシングの目的について, プロフェッショナルのミキシングエンジニアである山内 [35] は, 主役をはっきりさせることがミキシングの目的であるとしている. 加えて, 主役をはっきりさせるためには, マスキングを解消する必要があるとしている.

マスキングとは聴感特性の一種で, ある周波数の成分が, 時間方向, 周波数方向にある近い成分を聴こえなくしてしまう現象 [36] を指す. マスキングには, 周波数方向の同時マスキングと, 時間方向の継時マスキングとがある. つまり音信号は, 聴感特性において時間周波数平面上での拡がりを持っている.

マスキングについて, ミキシングのプロフェッショナルの間では, モノラルでのミキシングのとき, マスキングの影響がもっとも顕著となることが知られている [34].

マスキングを解消する処理は, 時間周波数平面上において成分の分布どうしの相互の干渉を解消することと等しい. この処理は, 時間周波数平面上で画像処理的にゲイン係数を生成することと一致する.

2.3.1　聴感特性

聴感特性について多くの研究が為されており, 様々な特徴が発見されている. この節では代表的なものについて紹介する.

まず, 人間の聴感は心理量である. 心理量は一般的に, 物理量である振幅やエネルギーなどとは対数関係にある [36]. これを, ウェーバー・フェヒナーの法則という.

人間の聴覚の感度は周波数帯域ごとに異なっている. このため, 同じエネルギーの異なる周波数の信号を聴いても同じ大きさには聴こえない. そこで, 周波数ごとに同じ大

表 2.1. チャンネルストリップに搭載される標準的な処理

調整する要素	処理	特徴
音量	フェーダー	$-\infty \sim +12$ dB のレンジを持つ
音色	イコライザ	ゲインのみ可変で, 3 もしくは 4 バンド構成
定位	パン	音量差ステレオとして動作

きさで聴こえるエネルギーについてまとめた聴感曲線が等ラウドネス曲線である. 関連して, 人間が知覚できる最低のレベルを最小可聴値と呼び, レベルが大きすぎるため人間が痛みを感じ始めるレベルを痛覚閾値 [37] と呼ぶ.

ある周波数の成分が, 時間方向, 周波数方向にある近い成分を聴こえなくしてしまう現象をマスキングと言う [36]. マスキングには, 周波数方向の同時マスキングと時間方向の継時マスキングとがある. 継時マスキングの効果は, 100 ms 程度の間は続く [38] とされている.

マスキングについて, ミキシングのプロフェッショナルの間では, モノラルでのミキシングのとき, マスキングの影響がもっとも顕著となる [34] ことが知られている. これを center masking と呼ぶ. center masking について, McQueen, Terhune(2011)[39] は, 脳での処理ではなく聴感特性によるものとして報告している.

2.3.2　よいミキシングとは

葛巻 [33] は, よいミキシングとして定位, 奥行き, 周波数分布がまんべんなく広がり, 隙間が無いものと述べている.

江夏 [30] は, ミキシングで最も重要視しなければならならないのが音の配置であり, 次に各パートの音量バランスが重要であると述べている.

これらの記述より本稿では, 音量のバランス, 定位のバランス, 音色のバランスの 3 点をミキシングのポイントとする. 第 2.4 節では, この 3 点について代表的なミキシングの処理を紹介する.

2.3.3　チャンネルストリップ

ミキシングのポイントである, 音量, 音色, 定位のバランスについて, 信号系統ごとに簡易な処理を一揃えにしたものをチャンネルストリップ [37] という. チャンネルストリップは 1970 年代後半ごろに登場し, ミキサーの基本の構成要素である.

チャンネルストリップに搭載される標準的な処理を, 表 2.1 にまとめる. チャンネルストリップの模式図を, 図 2.1 に示す.

チャンネルストリップに搭載されている処理では実施できない処理がある. 例えば, 位相差ステレオである. そこで, ミキサーは処理を追加できるように設計されている.

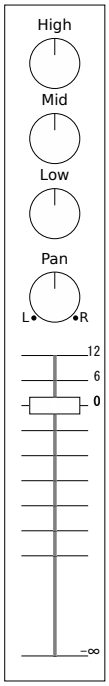


図 2.1. チャンネルストリップの模式図

プラグインという機能で, チャンネルストリップの処理と入出力との間に, 任意の処理を割り込んで接続できる.

2.3.4　音量の指標

2.3.4.1 実空間とデータ

実空間での音量とデータ上での音量は, 区別が必要である. 実空間での音量は絶対値である. 一方, データ上の音量は相対値である. データ形式でのダイナミックレンジにおいて, なにかしらの基準値との相対値でしかない. この基準値は, 再生機器, 再生環境, 聴取状況, 聴取者などの要素が総合されて決まる.

2.3.4.2 **物理量と心理量**　物理量と心理量としての音量は, 区別が必要である. 心理量としての音量は, 周波数特性がある.

まず, 人間の聴感は, 心理量である. 心理量は一般的に, 物理量である振幅やエネルギーなどとは, 対数関係にある [36]. これを, ウェーバー・フェヒナーの法則という,

人間の聴覚の感度は, 周波数帯域ごとに異なっている. このため, 同じエネルギーの異なる周波数の信号を聴いても, 同じ大きさには聴こえない. そこで, 周波数ごとに同じ大きさで聴こえるエネルギーについてまとめた聴感曲線が, 等ラウドネス曲線である. 関連して, 人間が知覚できる最低のレベルを最小可聴値という. 逆に, レベルが大きすぎるため人間が痛みを感じ始めるレベルを痛覚閾値 [37] という.

実空間での心理的な音量の値として, 騒音についての検討が主体である. A 特性, B 特性, C 特性の三種類があり, IEC 61672:2003 で規格化されている. それぞれ, 想定している騒音の聴取環境に違いがある.

2.3.4.3 VU/PPU メータ

ミキシング処理の主要な基準として, VU/PPU メータがある.

表 2.2 に, VU(volume unit) メータと PP(peak program) メータの代表的な規格の仕様を記す. VU メータの応答速度は 300 ms である. 一方, ツビスロッキは聴覚の時定数を 200 ms と推定している [40]. PPU メータでの応答速度は 2 種存在し, 5 ms と 10 ms とがある.

表 2.2. VU/PP メータの規格 ([41] に掲載の表に, [42] の情報を追加). 上 1 件が VU メータ, 下 3 件が PP メータの規格

規格番号	制定者	目盛り	基準レベル	attack time	decay time
ANSI C 16.5 IEC 60268-17	ベル研究所 CBS, NBC	−20 dB to +3 dB linear scale	0 VU = +4 dBu	99% in 300 ms	99% in 300 ms
IEC 60268-10 / I	Nordic(DIN)	−36, −30, −24, −18, −12, −6, TEST, +6, +9	TEST = 0 dBu	80% in 5 ms	20 dB in 1.5 s
IEC 60268-10 / IIa	BBC	1, 2, 3, 4, 5, 6, 7	4 = 0 dBu	80% in 10 ms	24 dB in 2.8 s
IEC 60268-10 / IIb	EBU	−12, −8, −4, TEST, +4, +8, +12	TEST = 0 dBu	80% in 10 ms	24 dB in 2.8 s

2.3.4.4 ラウドネスメータ EBU Mode

VU/PP メータに代わる音量のメータとして, LU メータ (ラウドネスメータ) が提案されている. LU メータの代表が, EBU が提案する EBU Mode(表 2.3)[43] である. 従来の PP メータに対応するのが, Momentary Loudness である. 従来の VU メータに対応するのが, Short-term Loudness である.

さらに EBU はこれらに加えて, Maximum True Peak Level を提案している. これらの値の算出はデジタル信号について行う. 算出する信号をアナログ信号としたときのピークレベルを推定する規格が, Maximum True Peak Level である.

EBU Mode は, 国内では ARIB TR-B32[44] として規定されている. 算出法は, 以下の 4 要素からなる.

- K 特性フィルタ (周波数重み付け)
- 二乗平均算出
- 各チャンネルのレベル重み付けと合算
- ゲーティング関数を適用したラウドネス値の算出
  - オーバーラップ法を用いた入力信号の分割 (ブロック化)
  - 2 段階の閾値によるゲーティング

算出法は, サンプリングレートを 48 kHz として規定されている.

K 特性フィルタについて述べる. K 特性フィルタは, 2 段の 2 次 IIR フィルタからなる. ここで 2 次 IIR フィルタ $H[z]$ を, 式 (2.1) とする.

$$H[z] = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{a_0 + a_1 z^{-1} + a_2 z^{-2}} \tag{2.1}$$

1 段目は頭部の影響による周波数特性である. 頭部形状を硬質球体に置き換えての特性を模擬する. 係数を, 表 2.4 に示す. 2 段目は RLB 特性 (修正 B 特性) のハイパスフィルタである. 騒音測定の B 特性を元にしている. 係数を, 表 2.5 に示す.

表 2.3. EBU Mode ([43] を参考に筆者が要約)

名称	略称	算出法	積分区間	更新間隔
Momentary Loudness	M	矩形窓	400 ms	-
Short-term Loudness	S	矩形窓	3 s	最低 10 Hz
Integrated Loudness	I	ITU-R BS.1770	“start” から “stop” まで	最低 1 Hz

表 2.4. K 特性フィルタ 第 1 段目の係数 [44] より

-	-	b_0	1.53512485958697
a_1	−1.69065929318241	b_1	−2.69169618940638
a_2	0.73248077421585	b_2	1.19839281085285

表 2.5. K 特性フィルタ 第 2 段目の係数 [44] より

-	-	b_0	1.0
a_1	−1.99004745483398	b_1	−2.0
a_2	0.99007225036621	b_2	1.0

2.3.4.5 等ラウドネス曲線

同じ物理量の純音であっても, 異なる周波数の信号での心理量としての音量は異なる. 心理量として同じ音量として感じる物理量をグラフ化したものが, 等ラウドネス曲線である.

ISO226 として, フレッチャー氏とマンソン氏による等ラウドネスレベル曲線が規格化されている. 4 kHz に, 特徴的なよく聞こえる周波数帯域が存在することが示されている.

IS0226:2003 として, 2003 年に新規格が制定された. 前規格の制定後の実験から, 誤差が発見されたことによる. 1 kHz 以下の低い周波数帯域では, 10 ∼ 15 dB の差 [45] がある.

2.3.5　ミキシングの方針

葛巻 [33] は, ミキシングの方針としてステージにある程度奥行きがあるコンサート・ホールを意識して行うと述べている. 聴感上での定位が, コンサート・ホール上での楽器の配置と一致するように調整する.

江夏 [30] は, 定位のバランスについてＶ字配置を勧めている. まず, 音の特性として低域の音ほど指向性を失い, 高域の音ほど指向性が強くなる. この特性に基づき, 低域は中央に配置するとバランスが良くなる, 高い周波数を持つ素材を左右に配置することで広がりや空間を作ると述べている. 加えて, 音像を左右に振り分けることで音の飽和感を改善できると述べている.

Izhaki[34] は center masking の特徴を用いて, マスキングの影響を確認すると述べている. プロフェッショナルは効率よくマスキングの影響を確認するために, ステレオでのミキシングの作業中であっても, モノラルでのミキシングも同時に用意する.

2.4　ミキシングの処理

この節では, ミキシングで施される様々な処理について紹介する.

2.4.1　処理とエフェクター

ミキシングの処理では, 狙った効果が得られるように, 様々な処理を適宜組み合わせて行う. ミキシングの処理の多くは, 音に様々な効果 (エフェクト) を与えるという意味で, エフェクター (Effecter, Effector, Effects unit) と呼ばれる [37]. 通常, Fader と Pan pot は, エフェクターには含まないが, 本論文ではすべて処理としてまとめて記述する.

エフェクターは, 信号の振幅処理, 信号の周波数処理, 信号の時間的処理の 3 系統に分類することができる [37].

2.3 節で述べたミキシングのポイントと, 主に関連する処理を, 表 2.6 にまとめる.

表 2.6. ミキシングのポイントに, 主に関連する処理

ミキシングのポイント	関連する処理
音量のバランス	Fader, 信号の振幅処理
定位のバランス	Pan pot, 信号の時間的処理
音色のバランス	信号の周波数処理

2.4.1.1　Fader

Fader(フェーダー) は, 相対的なゲインや音量を設定, 調整する機能である. レベルコントロール, ポテンショメーター, ゲインコントロールとも言われる [32].

ミキシングにおいて, 各素材を最適なレベルとすることは, 最も重要な要素である [30, 32]. この重要な役割を主に担うのが, フェーダーである.

2.4.1.2　Pan pot

Pan pot(パンポット) は, Panoramic Potentiometer の略語である [34, 37]. Panning(パンニング) とも呼ばれる.

ステレオの 2 チャンネルに, 音量差のある同一信号を流すことで, 2 つのスピーカ上の定位または移動を行う回路 [37] である. これにより, 人間の聴覚上での定位感を, 左右に振ることができる.

2.4.2　信号の振幅処理

信号の振幅処理に分類されるのが, Compressor や Limiter などの非線形処理である. ミキシングで用いられる代表的な非線形処理の特性を, 図 2.2 にまとめる.

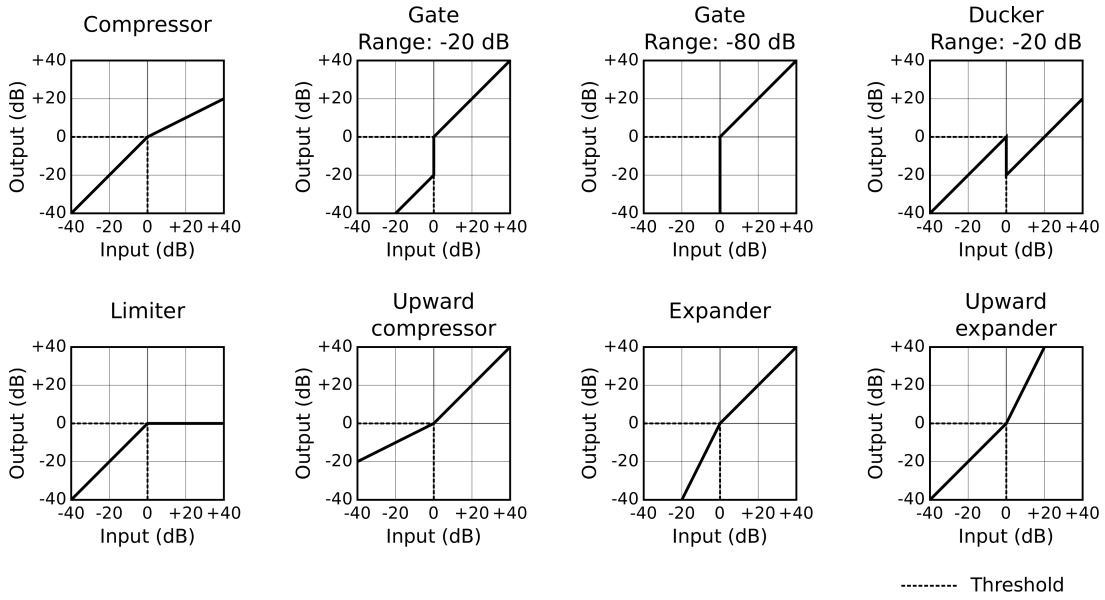


図 2.2. 代表的な非線形処理の特性. [34] に掲載の複数の図をもとに, 筆者が再描画

これらの処理には, 時間方向のパラメータも用意されていることが多い. Attack(アタック・タイム) と, Release(リリース・タイム) が代表的である. 前者が, 入力信号のレベルが閾値を上回ってから, 処理がかかりはじめる時間. 後者が, 入力信号のレベルが閾値を下回ってから, 処理がかかりおわる時間である [32].

非線形処理の特別な使用法として, side-chain(サイドチェイン) を用いたものがあげられる. サイドチェインは, 別の音信号の音量を用いて, 非線形処理を駆動させるというものである. サイドチェインは, Key-in という名でも用いられる.

振幅処理を使用する上での難点は, 駆動する周波数が特定されない点である. この難点の解決法として, 線形フィルターを組み合わせた処理が挙げられる. DeEsser, Multi Band Compressor が, その代表である.

2.4.2.1 Compressor

Compressor(コンプレッサー) は, 音量レベルを圧縮する [33] 処理である.

コンプレッサーで音量の凹凸をなくした信号は, 全体のレベルを上げることができる [33]. この特性により, 音圧を上げる処理にも用いられる.

また, アタック・タイムとリリース・タイムを調整することで, 音色や定位感を変化させる目的にも用いられる [33].

2.4.2.2 Limiter

Limiter(リミッター) は, いわばレベル管理に特化した Compressor(コンプレッサー) である [33]. コンプレッサーとの違いは, 閾値を超えた入力に対する出力との圧縮比に

ある. 一般的に, この圧縮比 (閾値を超えた入力の音量: 出力の音量) が 10: 1 以下をコンプレッサーといい, 8: 1〜∞: 1 を, リミッター [33, 37] という.

主な使用目的は, 機材の保護である. ピーク成分をリミット (制限) することで, 後段のレコーダーやスピーカーなどの機材に, 過大な音信号が入力されるのを防ぐ [33].

2.4.2.3 Expander

Expander(エキスパンダ) は, ダイナミックレンジを伸長する [37] 処理である. コンプレッサーとは逆の動作となる.

閾値以下の入力信号に対し, 出力を小さくなるように駆動する. これにより, 音の立ち上がりを鋭くし, メリハリのある音に変化させることができる [32].

2.4.2.4 Gate

Gate(ゲート) は, Expander の応用 [37] で, 閾値以下のレベルの信号をカットする. 無音部のノイズを除去する目的で使用される. その使用法から, Noise Gate(ノイズ・ゲート) と呼ぶこともある.

エレキギターなどのノイズが多い電気楽器, コンプレッサーなどを使用したことで発生したノイズに対して, 非常に大きな効果を発揮する [32].

2.4.2.5 Ducker

Ducker(ダッカー) は, 私たちにとって, 一般的な放送での処理として, 最も身近な非線形処理である. 例えば, DJ が話している時の, BGM の抑制に用いる [34].

ダッカーでは, サイドチェインを用いての使用が一般的である. サイドチェインの信号が閾値以上のとき, 入力信号を抑制する.

2.4.2.6 Multi Band Compressor

Multi Band Compressor(マルチバンドコンプレッサ) ーでは, まず入力信号を複数の周波数帯域に分け, それぞれの帯域について独立に非線形処理を行うものである. 分割する周波数帯域数は, 3 帯域が一般的である. この帯域数は, 処理能力としての充分さよりも, むしろ現実的に操作できるパラメータ数の制約を受けている.

マルチバンドコンプレッサーに対し, 単一の周波数帯域で処理をするコンプレッサーを, シングルバンドコンプレッサーともいう [33].

2.4.2.7 DeEsser

DeEsser(ディエッサー, De-essing) は, Limiter の一種で, 音声の子音を低減する目的で使用される [32]. 子音には高域成分が多く含まれる. そこで, 入力信号の高域成分が多くなったとき, コンプレッサーが動作することで, 音声を聞きやすくする [37]. ディエッサは, ダイナミック・シビランス・コントローラとも呼ぶ [37].

2.4.2.8 Exciter

Exciter(エキサイター) は, 入力信号の高域の偶数倍音あるいは奇数倍音を強調する処理である. 高域の倍音を強調することでメリハリをつけ, 音が全面に出るような効果を生み出す [37]. Exciter は, ハーモニクス系, 倍音系のエフェクトとも分類される [33].

倍音の発生には, 非線形処理を用いる. 非線形処理で発生した高域の倍音を, パラメトリックイコライザで任意に抽出し, 元信号に付与する.

Exciter を用いて低音楽器の音量感を補強できる. 低音楽器の基本周波数と倍音関係にある高い周波数成分を, 基本周波数成分と連動させて増幅する.　低い周波数帯域は, 音量が小さく聞こえる. このため, 音量を増幅するとダイナミックレンジの観点で難がある. 高い周波数帯域で音量増幅分を代替できることは, 限られたダイナミックレンジの中で調整を行う上で, 強力なツールである. これは, ミッシングファンダメンタル [46, 47] の効果であるとも解釈できる. 基本周波数成分がなくとも倍音列が存在するとき, 基本周波数成分が存在するように知覚される現象である.

2.4.3　信号の周波数処理

信号の周波数処理に分類される処理の代表的なものが, Equalizer である. 高次倍音を強調する Exciter も, この分類に含まれる.

2.4.3.1 Equalizer

Equalizer(EQ, イコライザー) は, ローパスフィルタ, ハイパスフィルタ, バンドパスフィルタの複合体である.

Equalizer を用いたレコーディングの流れを紹介する. 処理は二つの工程で行われる. まず楽曲全体について, おおまかに 各楽器のバランスを整える. このときに重要になるのが, 周波数領域の分離である. たとえば, スタジオ技術者用に書かれた専門書 [34] では, 図 2.3 のように周波数領域を分割して考えることが奨められている. これに準ずるように各領域毎に主役となる楽器を決定し, それ以外のパートには遠慮してもらうようにイコライザーによる調整が行われる. 図 2.4 が, バスドラムとベースそれぞれにかけるイコライザーの設定例である. それぞれのフィルターの特性が, 互い違いになっていることがみてとれる.

図 2.4 のように, 複数の音の組み合わせの中で, 周波数ごとに抑制と強調を行うイコライゼーションを, Mirrored equalization[34] と呼ぶ.

線形処理の問題点は, 音量によって効果が変化する点にある. これは, 人間の聴覚の感度が周波数によって異なる [36] ことによる. そこで楽曲の各所ごとにパラメータを時変とし, 微調整を行う.

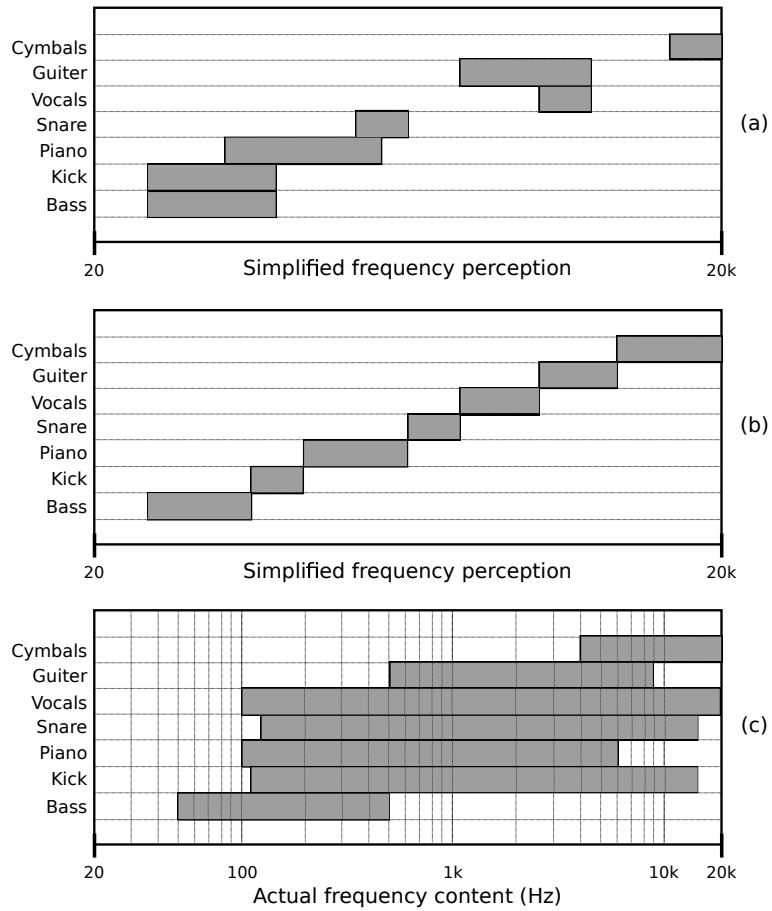
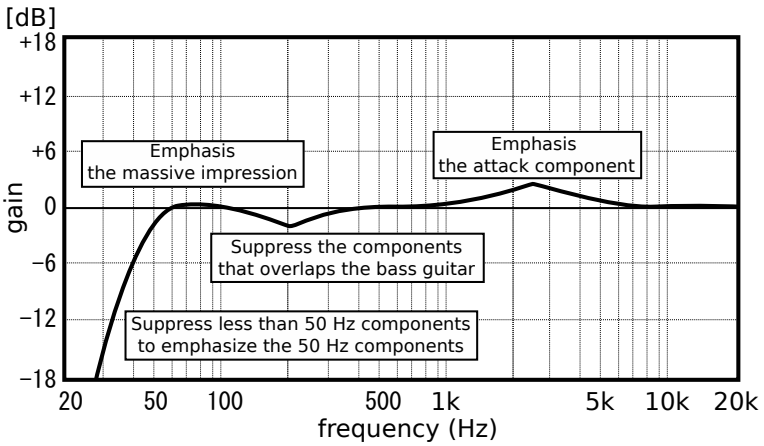
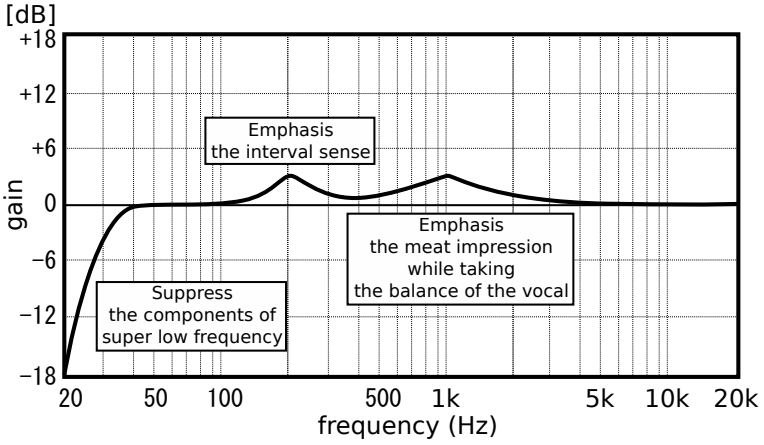


図 2.3. 周波数領域での楽器の振り分けの例. (a)：バランスの悪い振り分けの例. (b)：バランスのとれた振り分けの例. (c)：実際の楽器の周波数の分布. [34] に掲載の図を筆者が再描画



(a) Filter characteristic for bass drum



(b) Filter characteristic for bass guitar

図 2.4. イコライザーの設定例. デジタルフィルタの周波数特性. [48] に掲載の図を参考に, 筆者が再描画

2.4.4　信号の時間的処理

信号の時間的処理は, 空間系エフェクト [33] とも呼ばれる. その代表格は Delay(ディレイ) と Reverb(リバーブ) である. どちらも残響であることから, プロフェッショナルの現場では, まとめて Echo(エコー) と呼ぶことが多い [49]. サウンドレコーディング技術概論 [37] では, 山びこ効果のことをディレイと呼び, 風呂場やトンネル, 洞窟などで聴かれるような余韻をリバーブと呼ぶとしている.

ディレイを応用した, モジュレーション系エフェクトと呼ばれるエフェクトが存在する. コーラス, フランジャー, フェーザーである.

稲葉 [49] は, 時間的処理の各種類の効果について, リバーブが横と奥行き, ディレイが濁りの少ない奥行き, コーラスは横の広がりに効果があるとまとめている.

2.4.4.1 Delay

Delay(ディレイ) は, 遅延の名のとおり, 入力信号に対して遅らせた信号を付加する [33] 処理である.

入力信号に対する遅延量によって, ショート・ディレイ, ミディアム・ディレイ, ロング・ディレイなどと区別して呼ばれることがある. また, 遅延量として楽曲のテンポで与えるものをテンポ・ディレイと呼ぶ. それぞれの呼び名と一般的な遅延量の関係 [33] を, 表 2.7 に記す.

表 2.7. 呼び名と, 入力信号に対する遅延量の関係. [33] に記載の値をまとめた

呼び名	遅延量
ショート・ディレイ	20 ～ 50 msec
ミディアム・ディレイ	200 ～ 350 msec
ロング・ディレイ	400 ～ 500 msec 以上
テンポ・ディレイ	四分音符 1 つ

ショート・ディレイは, 入力信号に対し簡易ダブリング [33, 50] として用いられる. ダブリング (Doubling)[37, 33] とは, 演奏, 歌などを同じように 2 回録音しそれを同時に再生することによって太い音として聞かせる処理である. これは, 同じように演奏しても人間が行うために多少ずれることによる. ショート・ディレイが簡易ダブリングとして用いることができるのは, 入力信号と遅延信号の遅延時間を短く設定することで, 遅延信号が遅れて聞こえず, 入力信号と一体となって聞こえるためである [33].

2.4.4.2 Reverb

Reverb(リバーブ, Reverberation[37]) は, ディレイの応用 [33] で, 周期音やレベル, 周波数特性などが異なるディレイの出力を, 高密度に集合 [49] したものである. リバーブは, 自然の残響をシミュレートしているものが主である [49].

リバーブの処理として, 3 種の方法がある. 第一に, 実際の空間を代替として用いる方法である. スピーカによって空間に放射し再びマイクで集音する. 放射する空間として, 金属板や金属バネも用いる. 金属を用いることで, より省スペースで複雑な反射を得る狙いである. 特別に, 金属版を用いたものをプレートリバーブといい, 金属バネを用いたものをスプリングリバーブという. 第二に, 所望の空間のインパルス応答の畳み込みである. サンプリングリバーブ, IR リバーブと呼ぶ. 畳み込み演算であるため, 計算量の大きさが問題点である. 第三に, Schroeder Reverb[40, 51] である. 複数のディレイを直列および並列に組み合わせることで, 所望のインパルス応答を近似する.

リバーブは, 異なる種類のリバーブを複数組み合わせて用いるのが一般的である [33, 52]. 単体では, エフェクトの癖や品質の悪さが知覚されることがある. 組み合わせて用いることで, それらが互いに混ざり合い, 悪さが隠れ, 複雑で上質な残響感が得られる.

2.4.4.3 Phaser

Phaser(フェーザ, phase shifter[37]) は, 人工的に位相を変化させた音を原音と混ぜることによって, 直接音と間接音との干渉を模擬し, サウンドに深みを加える [37] ものである.

2.4.4.4 Flanger, Chorus

Flanger(フランジャー), Chorus(コーラス) は, ショートディレイを応用した [33] エフェクトである. 入力音を遅延時間を時変として原音に加算する. 代表的な効果として, ジェット機の上昇下降音のような響き [37] を作り出せる. その効果から, モジュレーション系エフェクトとも, 空間系エフェクトとも呼ばれる.

フランジャーとコーラスの違いは, 遅延時間の設定にある [13, 49]. フランジャーは LFO を用いて周期的に変動させる. これに対してコーラスはランダムに変動させる. また, フランジャーのディレイ・タイムはコーラスよりも短く設定される. フランジャーでは 0∼15 ms とし, コーラスでは 10∼25 ms とするのが一般的である.

2.4.5　ミキシングの処理例

本節では, 各種の処理を具体的な組み合わせた, ミキシングの処理例を紹介する. より単純な処理例から順に紹介し, 提案するスマートミキサーとの関係性を示す.

2.4.5.1 ミキシングの処理例 1: 単純加算（図 2.5(a)）

最も単純なミキシングでの処理は単純加算である.

この方法の利点は, まず処理が単純だということである. しかし, 入力する信号同士の周波数特性の相性によっては, マスキングにより音の聞き取りが妨害される場合がある.

2.4.5.2 ミキシングの処理例 2: 信号の切り替え（図 2.5(b)）

単純な足し算からの簡単な工夫として, 入力信号を切り替えて出力信号とする方法が挙げられる. 例えば, カーナビゲーションシステムにおいて案内音声を再生する際に, 既に BGM が再生されていたとしよう. このとき, 一旦 BGM の音量を切り, できた無音区間で音声を再生するというのが, この方法での動作である. 割り込む音声は確実に聞こえるが, BGM を楽しんでいたユーザに不快感を与える.

ここで, 切り替え処理を自動で行うために, 入力信号間に優先度の違いがあることを前提としていることに留意する. 優先度の高低に従って信号の切り替えを自動的に行うミキサーを, priority mixer[53] と呼ぶ.

2.4.5.3 ミキシングの処理例 3: 音量差の付与（図 2.5(c)）

例えばラジオ放送で, 話者によるナレーションと BGM とをミキシングする際に用いられる手法である. このミキシングでは, 実質的に入力信号どうしの間に優先度の差が存在する. ナレーションが主役であり, BGM は脇役である. そこで, 主役を十分に聞かせるために, 脇役の音量を主役に対して小さくする. つまり, 主役と脇役とに音量差を付与している.

このミキシングを実現するために, 加算器への入力の前に乗算器を接続する. 主役の入力信号の時間平均エネルギー, もしくはピーク値に応じて, 脇役への乗算器の係数を時変とする. このような処理は, 一般にダッカー（Ducker, 2.4.2 節）と呼ばれている. また, 処理対象の信号とは別の信号の情報によって, 処理を制御する方法を, サイドチェイン（side-chain）と呼んでいる.

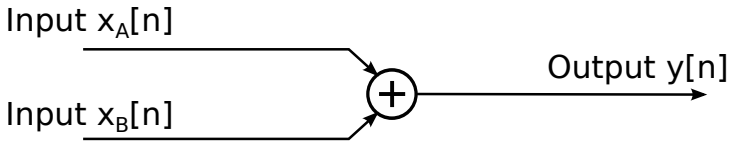
サイドチェインを用いた自動ミキサーとして, Dugan による Dugan mixer[54] がある. Dugan mixer では, 各入力信号について有音判定をし, 有音であると判定された信号数に応じて, 各入力信号への音量の上限値を時変に設定する.

ナレーションと BGM との最適な音量差について, 小森ら [18] は, ナレーションの重み付ラウドネスレベルが背景音より 9 phon 程度大きければ, 最適レベルと判断するとしている.

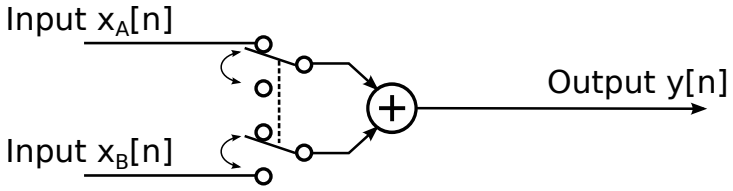
2.4.5.4 ミキシングの処理例 4: 非線形関数を用いた音量制御（図 2.5(d)）

入力信号どうしの音量差を調整する上で問題になるのが, 各々の入力信号が時間によってエネルギが変動することである. そこで, 音量差を与える乗算器の前に, 入力信号のダイナミックレンジを調整する非線形関数を追加することが行われる. ここで用いられる非線形関数は, 一般的にはコンプレッサ (2.4.2 節) である.

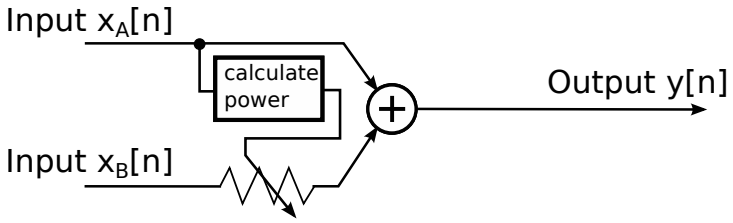
非線形関数での処理の問題点として, パラメータの設定によっては, 音質を著しく損なうことがあげられる. 非線形関数では, スペクトルの考慮なしに振幅を時間軸で変化させる. これによって, 波形そのものが書き換えられ, 同時にスペクトルが変化するた



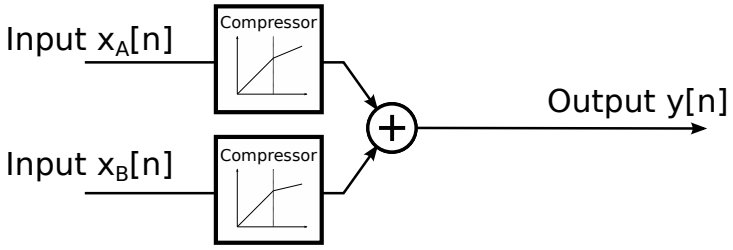
(a) 単純加算



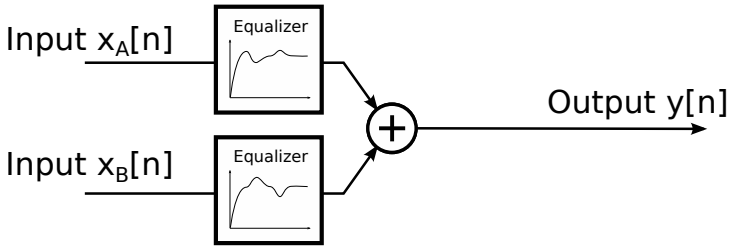
(b) 信号の切り替え



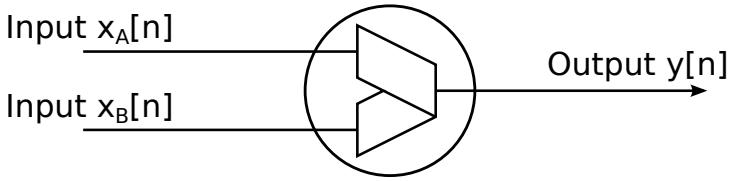
(c) 音量差の付与



(d) 非線形関数（コンプレッサ）を用いた音量制御



(e) 線形フィルタ（イコライザ）を用いた音色調整



(f) スマートミキサーへ

図 2.5. 単純加算からスマートミキサーへ

めである.

この問題点を解決する方法として, 周波数帯域ごとに異なるパラメータを設定するという方法がある. これを一般に, マルチバンドコンプレッサ（Multi Band Compressor, 2.4.2 節）と呼ぶ.

2.4.5.5 ミキシングの処理例 5: 線形フィルタを用いた音色調整（図 2.5(e)）

多くの音信号は, 固有のスペクトルの形状を持っており, 音色を決定づける特徴的な周波数帯域が存在する.

しかし, 数多くの信号を加算することで, 各入力信号の特徴的な周波数帯域どうしが被り, 音色が判別できなくなってしまう. 出力信号において各入力信号の特徴を生かすためには工夫が必要である.

そこで, 各入力信号のスペクトルの特徴が引き立て合うように, 線形フィルタ（イコライザ, Equalizer, 2.4.3 節）を用いて音色を調整する. この手法は, 2 信号にかける周波数特性が鏡に写したように逆になることから, Mirrored equalization[34] と呼ばれている.

2.4.5.6 スマートミキサーへ（図 2.5(f)）

処理例 4 で振幅への処理を, 処理例 5 で周波数領域での処理を紹介した. いずれの処理でも, 時区間によって処理の効果が一様でないという問題点がある. これは, 入力信号が非定常であることが原因である.

この問題点を解決するために, スマートミキサーでは振幅への処理では周波数帯域別に, 周波数領域での処理ではエネルギ別に, それぞれパラメータを調整することを考えた.

2.5　ミキシングの難しさ

ミキシングの難しさの要因として, 次の 2 点が考えられる.

第一に, エフェクターの効果を狙い通りに組み上げていくのが難しいことが挙げられる.

例えば, コンプレッサーを用いて音量を揃えたとき, 同時に音色にも変化が起きてしまう. その変化が意に反する場合は, イコライザーなどを用いて音色を修正する必要がある. 或いは, 種類の多い非線形処理の, 別の種類を試すことで解決に至るかもしれない. こうした試行錯誤を, 所望の処理が得られるまで続けることになる.

第二に, 聴感特性を考慮した処理を, 聴覚上の印象そのもので操作できる方法が存在しないことが挙げられる.

例えば, 同じ音信号であっても, 再生する音量が異なるとき, 音色が違って聴こえる. ある周波数の成分が, 時間方向, 周波数方向にある近い成分を聴こえなくしてしまう. これらは, 聴感特性によるものである.

こういった難しさを乗り越えるには, 長い年月をかけて蓄積された経験によって鍛えられた熟練の技を会得する必要がある.

これらの難しさを一挙に解決する方法として, 筆者らはスマートミキサーを提案した. スマートミキサーは, ミキシングを入力信号の時間周波数平面どうしの非線形な重ね合わせとして実施する. 時間周波数平面を導入することで, 音量と音色を分離よく処理することができる. 時間周波数平面は聴感特性を処理に反映させる基盤としても有用である. 聴感特性は時間周波数平面上で時間, 周波数方向に拡がりを持ち, 親和性が高い.

第 3 章
スマートミキサー

本章では, スマートミキサーについて述べる.

3.1 スマートミキサーとは

スマートミキサーとは, ミキサー内部に時間周波数平面による解析部を持ち, ミキシングを時間周波数平面の重ね合わせとして実施する音信号のミキサーである. スマートミキサーによって, 従来のミキシング法でプロフェッショナルが行なっていた高度のミキシングを, 手軽に実現できる. 加えて, 従来のミキシング法を包含しつつ, さらに緻密で自由な処理を提供できる.

スマートミキサーの基本構成を図 3.1 に示す. 2 本のデジタル音信号 $x_{\mathrm{A}}[n]$, $x_{\mathrm{B}}[n]$ を, スマートミキサーに入力する. ここで, n をサンプル番号とする. 入力信号は, まず周波数分解する. 本稿では, 周波数分解を FFT で行うこととする. 周波数分解は, 必要に応じてフィルタバンク, ウェーブレット変換などの異なる手法でも良い. 周波数分解後の i フレーム, k 帯域を, 時間周波数平面の座標 $[i,k]$ として示すこととする. この座標 $[i,k]$ を用いて, 入力信号 $x_{\mathrm{A}}[n]$, $x_{\mathrm{B}}[n]$ の時間周波数表現を, $X_{\mathrm{A}}[i,k]$, $X_{\mathrm{B}}[i,k]$ とする. 周波数分解に FFT を用いるため, $X_{\mathrm{A}}[i,k]$, $X_{\mathrm{B}}[i,k]$ は複素数値である. 入力信号の時間周波数表現 $X_{\mathrm{A}}[i,k]$, $X_{\mathrm{B}}[i,k]$ を, 非線形関数 $F(X_{\mathrm{A}}[i,k], X_{\mathrm{B}}[i,k])$ によって処理し, 出力信号の時間周波数平面 $Y[i,k]$ を算出する. 最後に, 出力信号の時間周波数平面 $Y[i,k]$ を IFFT し, 出力信号 $y[n]$ を得る.

スマートミキサーの特徴について, 従来ミキサーにおけるチャネルストリップとの比較で述べる. 図 3.2 は, 従来ミキサーの基本構成をチャネルストリップで示したものである. 従来ミキサーは, 複数のチャネルストリップを時間領域信号で結線したものと見

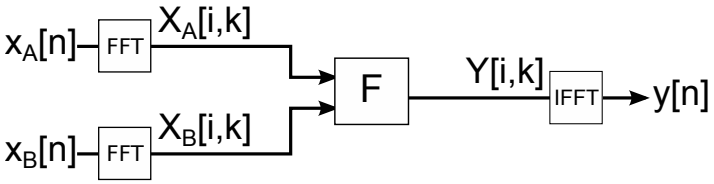


図 3.1. スマートミキサーの基本構成

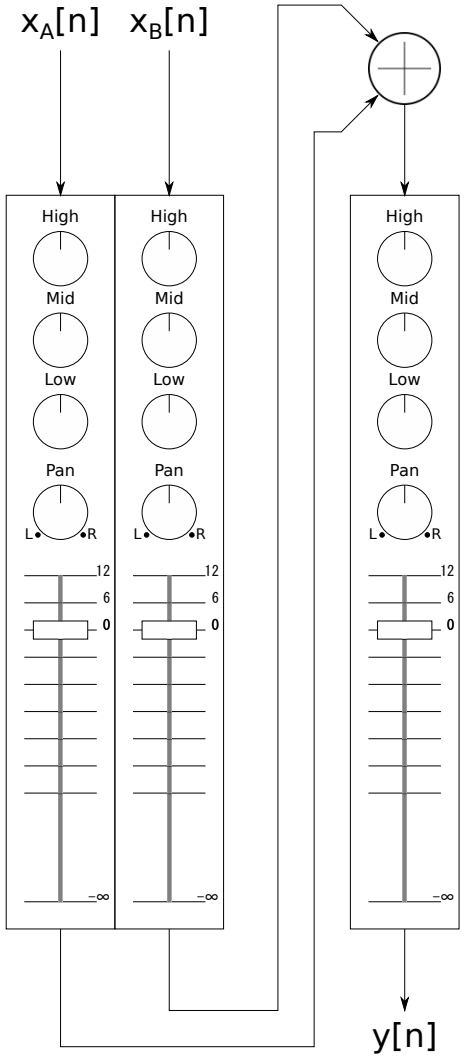


図 3.2. 従来ミキサーの模式図

なせる.

一方, スマートミキサーでは基本とする信号が時間周波数表現の 2 次元信号である. このため, 各エフェクト処理の相互作用を周波数帯域ごとに勘案でき, 実現できる処理が従来ミキサーに比べて広い. また, 従来ミキサーでは各チャネルストリップで独立した時間周波数解析及び波形再合成が行われている. スマートミキサーでは時間周波数解析を入力時に, 波形再合成を出力時に, それぞれ一元化できる. このため, スマートミ

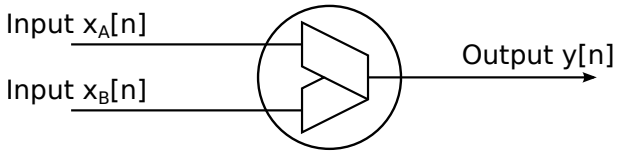


図 3.3. スマートミキサーを表す記号

キサーは従来ミキサーに比べて, ミキサー総体としての計算量を削減できる.

筆者らは, スマートミキサーを図 3.3 の記号で表している. この記号は, スマートミキサーの特徴を端的に表している. 記号を, 左側から右側に向けて説明していく. 2 本のデジタル音信号 $x_{\mathrm{A}}[n]$, $x_{\mathrm{B}}[n]$ を, スマートミキサーに入力する (図中左側, 円内に左から入る箇所). 入力されると, それらの音信号は, 時間周波数平面に展開される (図中左側, 横棒から縦棒に変化する箇所). 各入力信号の時間周波数平面 2 枚の非線形な重ね合わせを行う (図中中央, 2 つの平行四辺形が重なりあう箇所). 得られた 1 枚の時間周波数平面を, 時間領域信号に変換する (図中右側, 縦棒から横棒に変化する箇所). こうして, 所望の出力信号が得られる (図中右側, 円外に右へと出る箇所). これが, スマートミキサーの基本的な構造である.

3.2　スマートミキサーの研究目的

スマートミキサーの研究目的として 5 つ述べる. 第一に, スタジオ技術者としての経験のない個人に対し, プロフェッショナルが行う緻密なミキシングを, 手軽に実行できるシステムを提供することである. 第二に, スタジオ技術者に対し, 緻密で且つ自由な処理を実現できるミキシングシステムを提供することである. 第三に, 複数の音情報が同時に入力されるスピーカーシステムに対し, より多くの音情報を快適に発することができる, ミキシングアルゴリズムを提供することである. 第四に, ミキシングにおける聴感特性の知見を得ることである. 第五に, 時間周波数平面を用いた音信号処理の知見を得ることである.

3.3　スマートミキサーの実装

スマートミキサーの実装し得る形態を提案する. 実装は大きくハードウエアとソフトウエアとに分けることができる.

3.3.1　ハードウエアとしての実装

ハードウエアとしての実装の形態について述べる.

第一に, スマートミキサー装置である. 各種音響機器や楽器からの入力, 外部スピーカーシステムへの出力を, 汎用の端子で直接接続できるミキサー装置としてハードウエ

ア化する. スマートミキサー装置を従来ミキサーと接続できるようにし, 従来ミキサーの操作性をスマートミキサーの操作法の選択肢として取り入れることができると, よりスムーズにスマートミキサーを導入できる.

第二に, スマートミキサー SoC(System-on-a-chip) である. リアルタイムで動作するスマートミキサーの処理を SoC としてまとめることで, 各種機器の回路にスマートミキサーを導入できる.

3.3.2　ソフトウエアとしての実装

ソフトウエアとしての実装の形態について述べる.

第一に, スマートミキサープラグインである.　コンピューターでの音楽制作は, DAW(degital audio workstation) と呼ばれる, 音楽制作に関係する各種ツールを統合した GUI アプリケーションで行われることが一般的である. DAW 上でミキシングの処理の組み上げは, プラグイン (plugin) と呼ばれる拡張モジュールを追加することで行う. スマートミキサーのアルゴリズムをプラグインとして実装することで, 従来の DAW をスマートミキサーに改造できる.

第二に, スマートミキサーオーディオデバイスである. コンピューターでのリアルタイムの再生は, オーディオデバイスが行なっている. コンピューター上で起動する様々なアプリケーションが生成する音信号にスマートミキサーを用いるためには, オーディオデバイス自体がスマートミキサーとして動作する必要がある. オーディオデバイス自体がスマートミキサーとなることで, コンピュータの処理能力, 出力デバイスの設定に合わせたミキシングを行える.

第三に, クラウドスマートミキサーである. 動画投稿サイトへの作品投稿でのミキシングや IP 電話に対し, スマートミキサーの処理を行うサーバーを設置する. Web に接続されていることで, ユーザーの意見を集め, アルゴリズムに反映させることができる.

3.3.3　スマートミキサーによる社会貢献

スマートミキサー研究によって, 音の聴取環境をグレードアップできる. これによって, 音楽試聴やコンテンツ制作のみならず, 以下のような社会貢献ができる.

第一に, 音を用いた情報伝達の利用促進である. コンピュータ技術の発達を用いて, より多くの情報を素早くやりとりする技術が模索されている. スマートミキサーによって, 複数の音信号への処理の高度化と導入しやすさが実現できる.

第二に, 音による情報伝達の信頼性向上である. スマートミキサーの構成法を提案し検討することで, ミキシングにおける聴覚特性の知見が得られる. 発展して, 聞き逃しや勘違いのメカニズムの解明に貢献できる. これにより, 音による状況判断や情報伝達を, より確かなものにできる.

第三に, 事故災害対策である. 例えば, トンネル工事の現場でのスマートミキサーの活用である. 作業場内の環境にあわせて警報音の特性を変更し, 確実に作業員が気づく

ミキシングを実現したい．また, 複数の警報音を併用し, より多くの情報を作業員に提示できる．刻一刻と変化し, 全体像が把握しづらい現場では, 作業員自身のとっさの判断が大切だと思われる．音で判断の助けがしたい．

第四に, スマート家電への導入である．電話, Web などの家の外との音信号のやりとりは圧縮符号化されたデータを用いるのが主体である．圧縮符号化は, 符号化および復号の中で時間周波数表現を用いる．家の内外の音信号の整理を時間周波数表現を基盤として行うことで, 家全体としての処理の効率を向上することができる．

3.4　音声を埋もれさせない音信号混合法

本論文では, 音声を埋もれさせない音信号混合法を, 音声の内容の聴き取りを維持し, 音声と BGM の音質を維持する音信号混合法であると定義する．ここで, 音声や BGM の音質の壊れよりも, 音声の内容の聴き取りを優先することとする．

音声と BGM を入力とするミキシングは, 音楽製作やコンテンツ制作に留まらず, 日常生活の様々な場面で活躍している．例えば, カーオーディオにおけるナビゲーション音声と BGM がある．従来では, ナビゲーション音声が再生される時区間で, BGM の音量を減衰させていた．音量の増減が頻繁に起きると, 不快感に繋がる．音声を埋もれさせない音信号混合法の検討によって, 音響信号を活用できる可能性が広がる．

本論文では, 音声を埋もれさせない音信号混合法のスマートミキサーによる構成法として 3 手法を提案する．第一に, 調波構造に着目した手法であり, 提案法 A として第 4 章で述べる．第二に, フォルマント構造に着目した手法であり, 提案法 B として第 5 章で述べる．第三に, フォルマント構造と歯擦音に着目した手法であり, 提案法 C として第 6 章で述べる．

全ての提案法で共通する設定として, 入力信号に優先度を導入する．これは, 主役をはっきりさせる [35] という, プロフェッショナルのミキシングエンジニアの方針とも一致する．音声と BGM とのミキシングでは, 音声が主役である．従って, 音声は優先度が高い入力信号, BGM は優先度が低い入力信号とする．

第 4 章
音量関係と調波構造に着目した検討

本章では, 調波構造に着目した手法を提案法 A として提案する. 提案法 A は, 音声を埋もれさせない音信号混合法を実現する手法であり, 音声の時間周波数平面上のパターンを形作る調波構造を強調することで, 目的を達成しようとするものである.

まず, 4.1 節に調波構造を着目するに至った検討についてまとめる. 次に, 4.2 節に 4.1 節で得た知見を調波構造に着目した手法に置き換えた検討についてまとめる. 続いて, 4.3 節で 4.2 節で得た知見を反映した提案法 A を提案する.

4.1 音量関係に着目した手法

音声を埋もれさせない音信号混合法の第一の手法として, 入力信号の時間周波数平面上の音量関係に着目した手法を検討する. 入力信号の時間周波数平面上の音量関係に基づいて振幅と位相をそれぞれ変化させ, 処理の効果を確かめる. はじめに, 第 4.1.1 節で振幅を操作する手法を局所抑制法として検討する. 次に, 第 4.1.2 節で位相を操作する手法を局所同期法として検討する. 続いて, 第 4.1.3 節で局所抑制法と局所同期法を連携し, 効果を両立させる手法を検討する.

聴感上の埋もれは, 両信号間での音量差によって生じる. 優先度の高い入力信号よりも, 優先度の低い入力信号の音量が一定以上大きければ, 優先度の高い入力信号は聴きとれない. 従って, 優先度の高い入力信号を埋もれさせないためには, 優先度の低い入力信号の音量を下げればよい.

優先度の低い入力信号への最も単純な抑制は, 単に音量を下げることである. この処理は, 時間周波数平面での処理として見たとき, 全領域で音量を抑制することである. しかし, 全領域で音量を抑制する必要はない. 聴感特性として, 各周波数成分の知覚に

影響を及ぼす成分は, 近傍の周波数帯域に限られることが知られている. 優先度の高い入力信号の周波数成分の近傍以外の領域での抑制は, 埋もれを解消する役割は果たしていないはずである. この余分な抑制を廃することで, 優先度の高い入力信号の埋もれの解消と優先度の低い入力信号の過度な音量減衰感の低減を両立できる.

時間周波数平面を用いない従来法の問題点を, 聴感上で不必要な時間周波数平面上での領域へも抑制する点にあると考える. そこで, 聴感上で必要な時間周波数平面上での領域のみを局所的に抑制, 位相同期を行う. 時間周波数平面の各点の振幅と位相をそれぞれ変化させる手法を, 局所抑制法と局所同期法とする. 時間周波数平面上での処理領域を必要最小限にとどめることで, 処理に伴う音量減衰を最小限に留めることを狙う.

4.1.1　音量関係に着目した局所抑制法

本節では, 時間周波数平面上での振幅処理として, 2 手法の局所抑制法を検討する. 音信号を用いてパラメータ調整し, 音声を埋もれさせない音信号混合法としての効果の確認と, 効果的な処理の条件を考察する.

局所抑制法は, 主役をはっきりさせる [35] という, プロフェッショナルのミキシングエンジニアの方針に倣った方法である. 主役である音声以外を抑制することで, 音声が埋もれないミキシングの実現を目指す.

4.1.1.1　一般形

局所抑制法の一般形を, 図 4.1 に示す. 入力信号を, 優先度の高いものから順に $x_{\mathrm{A}}[n]$, $x_{\mathrm{B}}[n]$ とし, 出力信号を $y[n]$ とする. ここで, n は時刻 (サンプル番号) である. $X_{\mathrm{A}}[i,k]$, $X_{\mathrm{B}}[i,k]$, $Y[i,k]$ は, 各信号の時間周波数平面上の位置 (i,k) での短時間フーリエ変換値である. ここで, i は時間フレーム, k は周波数ビン番号である.

$X_{\mathrm{A}}[i,k]$, $X_{\mathrm{B}}[i,k]$ のパワー $|X_{\mathrm{A}}[i,k]|^2$, $|X_{\mathrm{B}}[i,k]|^2$ を dB 表記したものを, $L_{\mathrm{A}}[i,k]$, $L_{\mathrm{B}}[i,k]$ とする. $L_{\mathrm{A}}[i,k]$, $L_{\mathrm{B}}[i,k]$ から非線形関数 $F(L_{\mathrm{A}}[i,k], L_{\mathrm{B}}[i,k])$ によって, $X_{\mathrm{B}}[i,k]$ に対するゲイン係数 $W_{\mathrm{B}}[i,k]$ を算出する. 出力 $Y[i,k]$ を, $Y[i,k] = X_{\mathrm{A}}[i,k] + W_{\mathrm{B}}[i,k]X_{\mathrm{B}}[i,k]$ で決定する.

なお, $F()$ は周波数帯域ごとに異なる関数とする. 従って, 本来は $F(L_{\mathrm{A}}[i,k], L_{\mathrm{B}}[i,k], k)$ と表記すべきであるが, 紛れのある時を除いて $F(L_{\mathrm{A}}[i,k], L_{\mathrm{B}}[i,k])$ と表記する.

4.1.1.2 手法 1　音量関係に基づいて抑制量を決定する方法

$F()$ について議論する. $F()$ の決定は, $L_{\mathrm{A}}[i,k]$ と $L_{\mathrm{B}}[i,k]$ が作る 2 次元平面上の各点で $W_{\mathrm{B}}[i,k]$ を決定することに相当する. 本手法では, 図 4.2 に示すように, 2 次元平面上に 2 つの長方形を配置し, それらを基準に W_{B} を決定することとした. まず, 内側の長方形について l_{A2}, l_{A3}, l_{B2}, l_{B3} で位置を決定し, 長方形内の抑制量をパラメータ $w_{\max}$ (dB) とする. 次に, 外側の長方形について l_{A1}, l_{A4}, l_{B1}, l_{B4} で位置を決定し, 長方形外の抑制

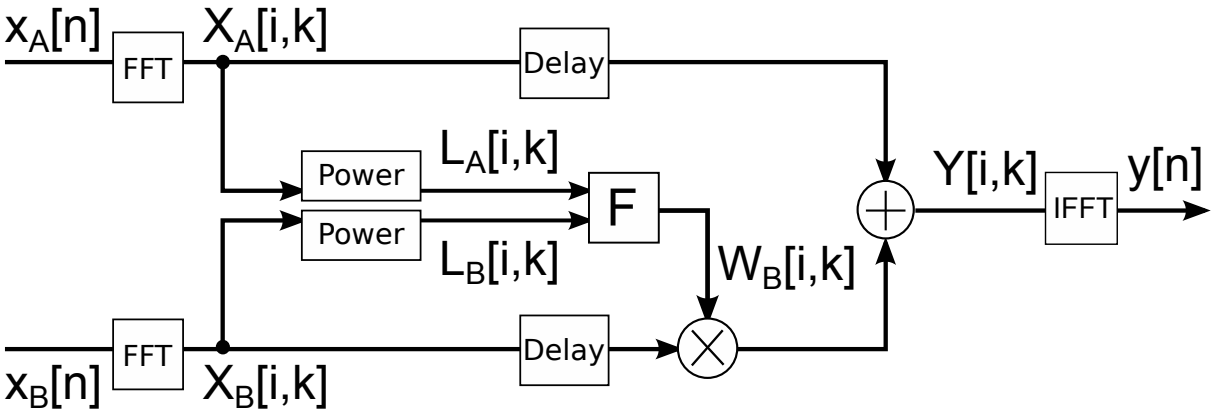


図 4.1. 局所抑制法のブロック図

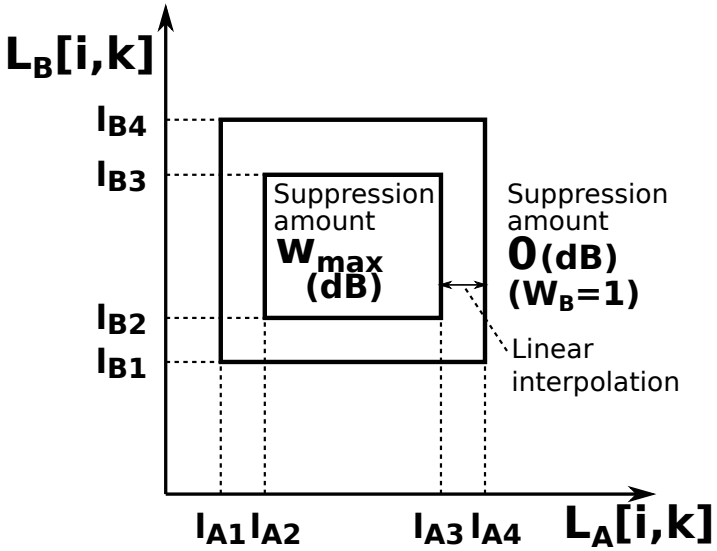


図 4.2. 非線形関数 $F(L_{\mathrm{A}}[i,k], L_{\mathrm{B}}[i,k])$ の特性

量を 0 dB とする. 続いて, 2 つの長方形の内側の領域について, 式 (4.1∼4.3) のように線形補間する.

表 4.1. 8 周波数帯域の設定

Band	Frequency [Hz]	Band	Frequency [Hz]
0	0 - 50	4	500 - 1250
1	50 - 125	5	1250 - 2500
2	125 - 250	6	2500 - 5000
3	250 - 500	7	5000 - 22050

$$W_{\mathrm{B}}[i,k] = 10^{\max(w_{\mathrm{A}},w_{\mathrm{B}})/20} \tag{4.1}$$

$$w_{\mathrm{A}} = \begin{cases} \dfrac{L_{\mathrm{A}}[i,k]-l_{\mathrm{A1}}}{l_{\mathrm{A2}}-l_{\mathrm{A1}}}\cdot w_{\max} & l_{\mathrm{A1}} \le L_{\mathrm{A}}[i,k] < l_{\mathrm{A2}} \\ w_{\max} & l_{\mathrm{A2}} \le L_{\mathrm{A}}[i,k] < l_{\mathrm{A3}} \\ \dfrac{l_{\mathrm{A4}}-L_{\mathrm{A}}[i,k]}{l_{\mathrm{A4}}-l_{\mathrm{A3}}}\cdot w_{\max} & l_{\mathrm{A3}} \le L_{\mathrm{A}}[i,k] < l_{\mathrm{A4}} \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

$$w_{\mathrm{B}} = \begin{cases} \dfrac{L_{\mathrm{B}}[i,k]-l_{\mathrm{B1}}}{l_{\mathrm{B2}}-l_{\mathrm{B1}}}\cdot w_{\max} & l_{\mathrm{B1}} \le L_{\mathrm{B}}[i,k] < l_{\mathrm{B2}} \\ w_{\max} & l_{\mathrm{B2}} \le L_{\mathrm{B}}[i,k] < l_{\mathrm{B3}} \\ \dfrac{l_{\mathrm{B4}}-L_{\mathrm{B}}[i,k]}{l_{\mathrm{B4}}-l_{\mathrm{B3}}}\cdot w_{\max} & l_{\mathrm{B3}} \le L_{\mathrm{B}}[i,k] < l_{\mathrm{B4}} \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

次に, 周波数特性について説明する. 9 個のパラメータを各 k について設定するのは自由度が高すぎて扱いにくいため, 表 4.1 に示す 8 個の帯域にまとめた.

予備的に行った聴き取り実験では, 帯域 0 と帯域 7 では $W_{\mathrm{B}} = 1$ とするのが最適であった. そこで以後, 帯域 1 から 6 の 6 帯域 (50 - 5000 Hz) のみを F による抑制の対象とする.

実験素材として, 音声と BGM それぞれ二つのデータを用意した. 音声には, 話速バリエーション型音声データベース:SRV-DB[55] より, 男性と女性の発話データを一文ずつ用いた. 以後, 男性のデータを PM00, 女性のデータを PF00 と表す. BGM には, クラシック (BGM1) とジャズ (BGM2) を選択した. これら 4 つの素材について, PM00 と BGM1, PM00 と BGM2, PF00 と BGM1, PF00 と BGM2 の, 4 つの組み合わせのミキシングを試した.

まず, 各素材の有音部の二乗平均値について, BGM が音声より 12 dB 大きくなるようにレベルを調整した. このとき単純加算では, すべての組み合わせで音声を聞き取ることができない. ここで, 4 つの実験素材の有音部は図 4.3 のように指定した. (a), (b) の波形の上部のひらがなは, 該当する箇所での発話内容である.

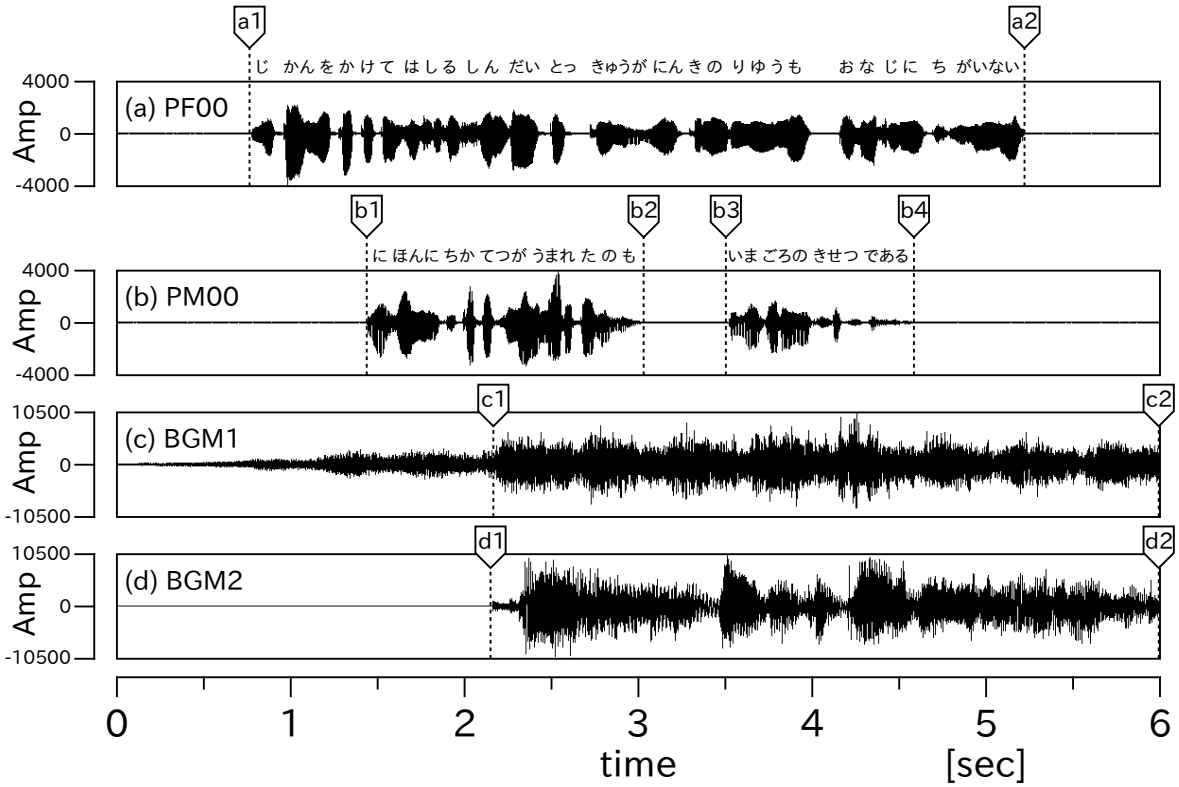


図 4.3. 実験素材の波形

表 4.2. スマートミキサーの実験諸元

サンプリング周波数	F_{s}	44.1 kHz
入力信号 A	$x_{\mathrm{A}}[n]$	音声 (PM00:日本に地下鉄が生まれたのもいまごろの季節である PF00:時間をかけて走る寝台特急が人気の理由も同じに違いない)
入力信号 B	$x_{\mathrm{B}}[n]$	BGM (BGM1:ベートーベン作曲交響曲第 9 番 BGM2:A 列車で行こう)
FFT 点数	N_{FFT}	4096 点
解析時窓関数	$H[n]$	ハニング窓形 2048 点
合成時窓関数	$G[n]$	ハニング窓形 1024 点
フレームシフト	N_{SFT}	512 点

良好な結果が得られた処理を, 帯域 3 を例にとって紹介する.
まず, 図 4.4 に帯域 3 において $L_{\mathrm{A}}[i,k]$ を横軸に, $L_{\mathrm{B}}[i,k]$ を縦軸にとった散布図を示

す. 図 4.4 中の斜めの破線は, $L_{\mathrm{A}}[i,k]$ と $L_{\mathrm{B}}[i,k]$ のエネルギ差が 0 dB のラインである.

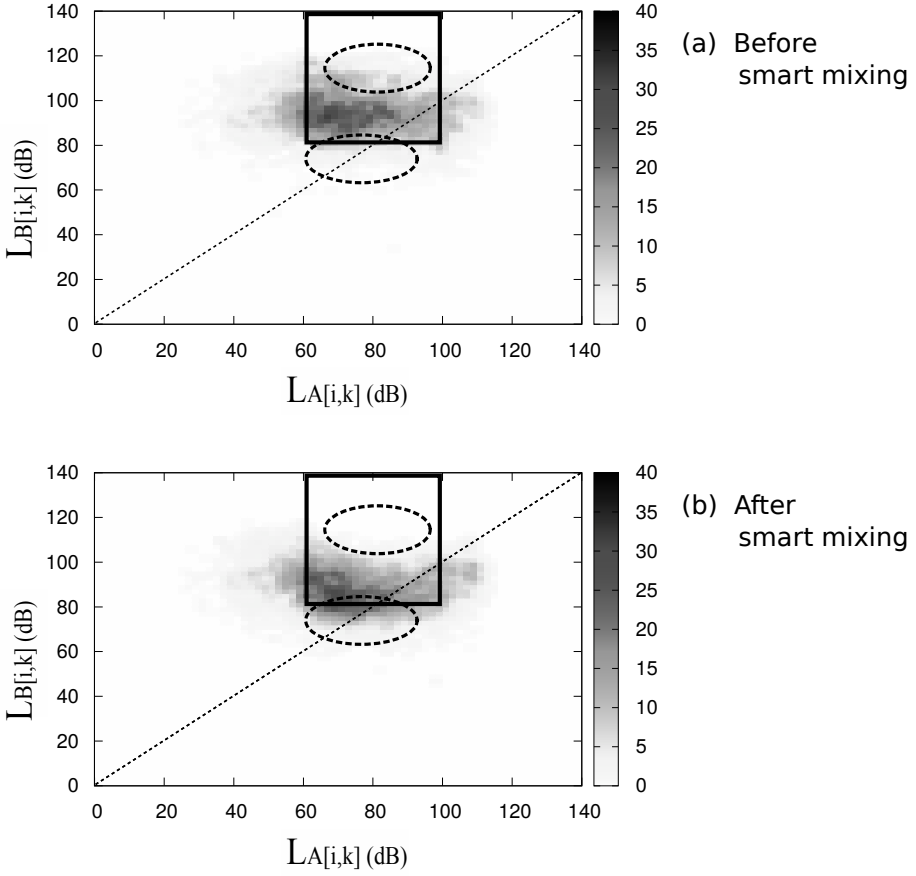


図 4.4. 帯域 3(250 - 500 [Hz]) での, $L_{\mathrm{A}}[i,k]$ と $L_{\mathrm{B}}[i,k]$ の散布図

このラインより上部では, $L_{\mathrm{A}}[i,k]$ が $L_{\mathrm{B}}[i,k]$ よりも小さい.

図 4.4 中の四角枠内の時間周波数分布に対して, $F(L_{\mathrm{A}}[i,k], L_{\mathrm{B}}[i,k])$ を w_{Max} を 18 dB として駆動させた. この処理による四角枠内の時間周波数分布の図 4.4 中での変化は, 下方向に最大 18 dB 下方への移動である. 例えば, 図 4.4 中の 2 つの破線楕円枠で変化がみてとれる. 上方にある楕円内での密度は, 処理前と処理後を比較して減少している. 対して, 下方にある枠内では, 処理前と処理後で, 密度の上昇がみられる. このように, データの集中する位置が図中の下方に移動している. これは, 時間周波数平面上の位置 (i,k) で, 優先度の高い信号のエネルギが優先度の低い信号のエネルギに対して, 小さい値となっている箇所が減っていることを意味する.

この設定で処理をした結果, 他の設定に対して比較的良好な結果が得られた. 単純加算では聞き取りができなかった音声が聞き取れるようになったともに, BGM の劣化が比較的少量となった. ただし, BGM の音色の印象は, 元信号に対して変化が感じられた.

図 4.5 に, 二つの入力信号, 単純加算, 処理結果, 抑制後の優先度の低い信号の時間周波数平面のうち, 0 から 3 kHz までの範囲を示す. 単純加算では, 音声の線スペクトル構造の特徴が際立っておらず, ほとんど音楽に埋もれている. これに対して, 処理後の

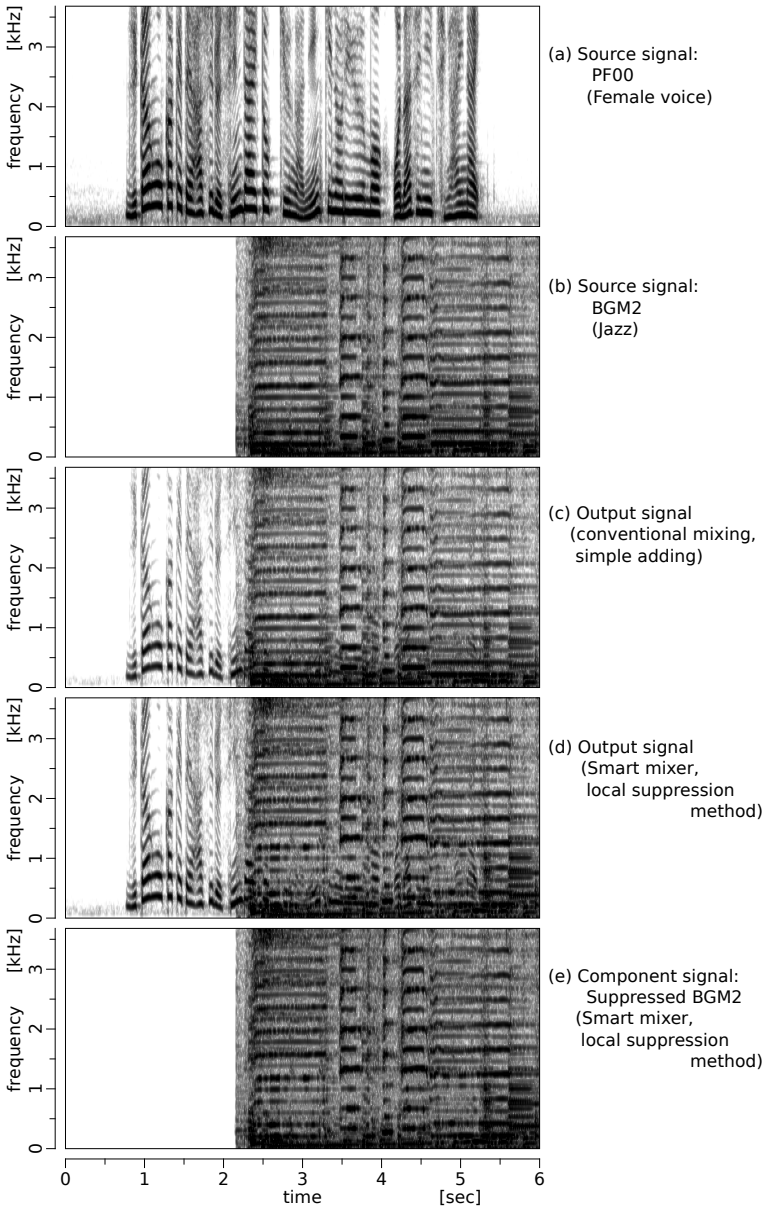


図 4.5. 処理前後の信号の時間周波数平面

信号では音声の線スペクトル構造がみられるとともに, BGM の線スペクトル構造が削られている. BGM の時間周波数分布を, 音声のスペクトル構造を崩さないように変更することができたといえる.

本手法についてまとめる. 特定の音声と BGM のミキシングについて, 最適な抑制量が得られる条件について, 次の 4 つの知見を得た. 第 1 に, 時間周波数平面上で, 優先度の高い信号と低い信号との音量差に着目するべきである. 第 2 に, 周波数毎のエネルギの分布に即して, 抑制量を調整するべきである. 第 3 に, 50 Hz より低い帯域, 5,000 Hz より高い帯域では, 抑制を行うべきではない. 第 4 に, 両入力信号について, 時間周波数

平面上でエネルギが一定値より大きい箇所では, 抑制を行うべきではない.

4.1.1.3 手法 2　音量差に着目した手法

4.1.1.2 節で述べた手法 1 で得られた知見を元に, 音量差に着目した手法を検討する. 本手法では, ゲイン係数 $W_{\mathrm{B}}[i,k]$ を式 4.4,4.5 とする.

$$W_{\mathrm{B}}[i,k] = \begin{cases} 10^{(\delta-\Delta[k])/20} & (\delta < \Delta[k], \\ & l_{\mathrm{a}}[k] \leq L_{\mathrm{A}}[i,k] < l_{\mathrm{A}}[k], \\ & L_{\mathrm{B}}[i,k] < l_{\mathrm{B}}[k]) \\ 1 & (\text{otherwise}) \end{cases} \tag{4.4}$$

$$\delta = L_{\mathrm{A}}[i,k] - L_{\mathrm{B}}[i,k] \tag{4.5}$$

この非線形関数では, $L_{\mathrm{A}}[i,k]$ と $L_{\mathrm{B}}[i,k]$ の音量差 δ が, 閾値 $\Delta[k]$ よりも大きいとき, 処理後の音量差が閾値 $\Delta[k]$ となるよう, BGM を抑制する. ただし, $L_{\mathrm{A}}[i,k]$ が閾値 $l_{\mathrm{a}}[k]$ よりも小さいとき, 閾値 $l_{\mathrm{A}}[k]$ よりも大きいとき, $L_{\mathrm{B}}[i,k]$ が閾値 $l_{\mathrm{B}}[k]$ よりも大きいときは 抑制を行わない. 閾値 $l_{\mathrm{a}}[k]$, $l_{\mathrm{A}}[k]$, $l_{\mathrm{B}}[k]$, $\Delta[k]$ は, 周波数毎に調整する. 検討は, パラメータを共通とする周波数帯域を, 表 4.3 に示す 8 帯域とし, このうち, 帯域 1 から 6 までの 6 帯域 (50 - 5000 Hz) について行う. 従って, 50 Hz より低い帯域, 5,000 Hz より高い帯域では, 被優先音に対して抑制を行わない. これより, パラメータは計 24 個となる.

表 4.3. 周波数帯域の設定

Band	Frequency [Hz]			Band	Frequency [Hz]		
0	0	-	50	4	500	-	1250
1	50	-	125	5	1250	-	2500
2	125	-	250	6	2500	-	5000
3	250	-	500	7	5000	-	22050

この非線形関数の各帯域での特性は, 図 4.6 と表せる. 図中の実線枠内に分布する入力信号のエネルギ組について抑制する. 例えば, 入力時に図中の α に位置する成分は, 抑制処理後に斜線上の β へと移動するように抑制する. この時, 抑制量は α-β 間の矢印の長さに相当する. 同様に, 入力時に α' に位置する成分は, 抑制処理後に斜線上の β' へと移動するように抑制する.

非線形関数のパラメータ, $l_{\mathrm{a}}[k]$, $l_{\mathrm{A}}[k]$, $l_{\mathrm{B}}[k]$, $\Delta[k]$ を変えた処理を行った中から, 良好な結果が得られた処理例を紹介する.

入力信号は, 音声には SRV-DB[55] の「発話のプロフェッショナルによる編集手帳（読売新聞）の読み上げ」のデータを, 音楽には「RWC 研究用音楽データベース：ポピュラー音楽 [56]」のデータを用いた. 表 4.4 に示す諸元で処理を行った.

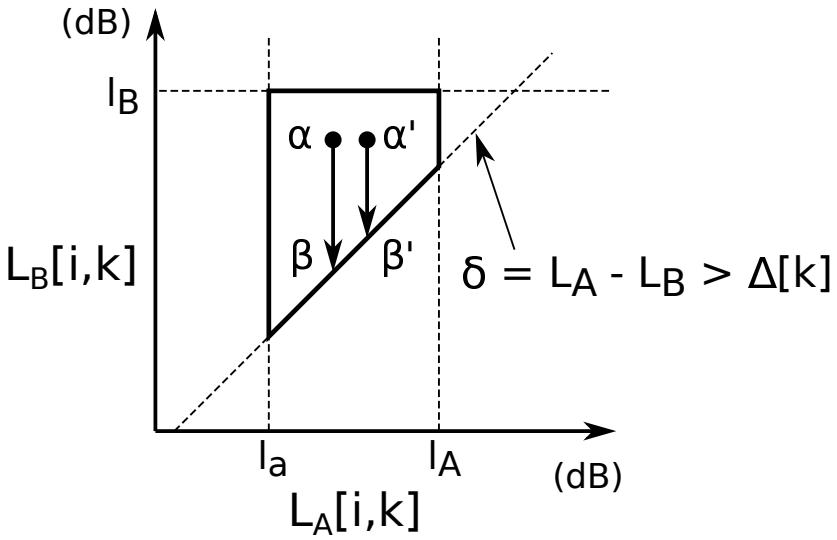


図 4.6. 局所抑制法 (手法 2) の非線形関数の特性

表 4.4. スマートミキサーの実験諸元

サンプリング周波数	F_{s}	44.1 kHz
FFT 点数	N_{FFT}	4096 点
解析時窓関数	$H[n]$	ハニング窓形 2048 点
合成時窓関数	$G[n]$	ハニング窓形 1024 点
フレームシフト	N_{SFT}	512 点

ここでは, 帯域 3 を例にとり実データを元に考察を行う.

まず, 図 4.7 に, $L_{\mathrm{A}}[i,k]$ を横軸に, $L_{\mathrm{B}}[i,k]$ を縦軸にとった散布図を示す. 図 4.7 中の二本の右上がりの破線は, $L_{\mathrm{A}}[i,k]$ と $L_{\mathrm{B}}[i,k]$ の音量差 δ の等高線である. 上側の破線上では 0 dB 差であり, 下側の破線上では 3 dB 差である. 第 3 帯域では閾値 $\Delta[k]$ を, 下側の破線の 3 dB とした. このパラメータ設定のとき, 太線内の成分がこの $\Delta[k]$ の破線に向けて真下に移動する. この移動量が, BGM の時間周波数成分へのゲイン係数 $W_{\mathrm{B}}[i,k]$ である. ここで, 散布図中で $L_{\mathrm{A}}[i,k]$ の最も大きい領域を太線が枠外とし, 抑制の対象外としている点が興味深い. この領域は $X_{\mathrm{A}}[i,k]$ では調波構造のピークに相当し, 太線が枠内とする領域はピークの周縁部となる.

図 4.7 中の 2 つの破線楕円 α, β の領域に着目する. α には, 処理前では多く分布していたが, 処理後では減少している. 対して β では, 処理前に比べ処理後で, 成分が増加している. 処理後では β の領域に値が集中するべきであるが, 実際に集中しているのは β の上部である. また, 本来は変化が起きない太線枠外でも, 処理による値の変化が出ている. これらの原理と動作のずれは, 処理後の合成窓関数による時間方向への処理の滲みが原因だと考えている.

図 4.8 に, 二つの入力信号, 単純加算, 局所抑制法での処理結果の時間周波数平面の

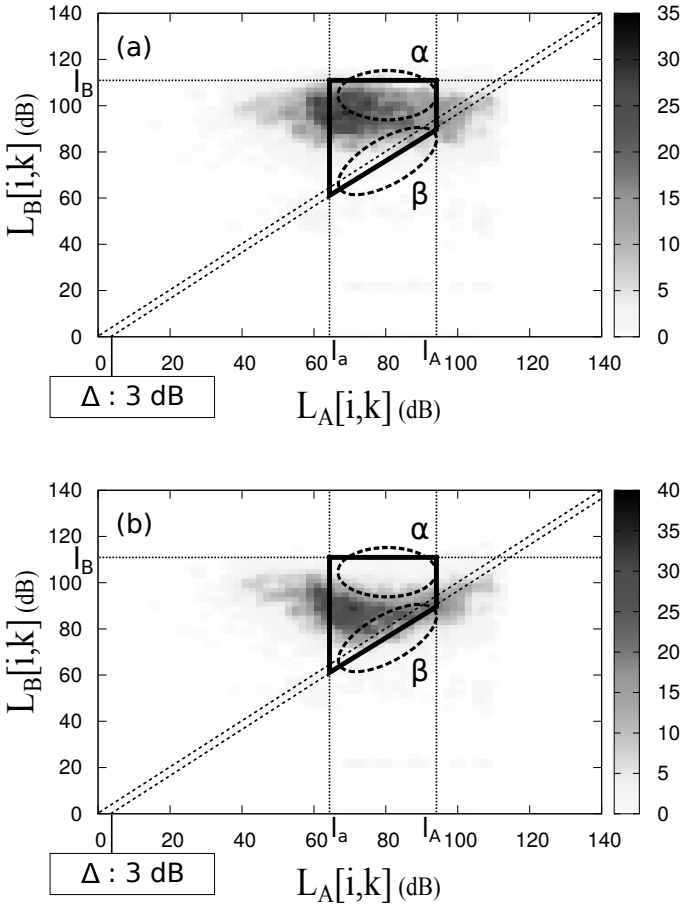


図 4.7. 帯域 3 における, $L_{\mathrm{A}}[i,k]$ と $L_{\mathrm{B}}[i,k]$ の散布図上での非線形関数の動作　(a) 処理前 (b) 処理後

うち, 特徴的な箇所を示す. 単純加算 (図 4.8(c)) では, 音声の線スペクトル構造の特徴が際立っておらず, ほとんど音楽に埋もれている. 対して, 処理後の信号 (図 4.8(d)) では, 音声の線スペクトル構造がみられるとともに, 音声の調波構造のピークの周縁部で, BGM の線スペクトル構造が削られている様子がみられる. BGM の時間周波数分布を, 音声のスペクトル構造を崩さないように, 変更することができたといえる.

4.1.2　音量関係に着目した局所同期法

本節では, 時間周波数平面上での位相処理である, 局所同期法の構成法を検討する. 実際の音信号でパラメータ調整し, 効果を確認すると共に, 音声を埋もれさせない音信号混合法で効果的な処理の条件を考察する.

局所同期法の概念について述べる.　2 つの入力信号の時間周波数平面上の 1 ビン, $X_{\mathrm{A}}[i,k]$ と $X_{\mathrm{B}}[i,k]$ を加算したとき, 出力信号 $Y[i,k]$ のパワーは, 入力信号成分のパワー $|X_{\mathrm{A}}[i,k]|^2$, $|X_{\mathrm{B}}[i,k]|^2$ の単純な足し算にはならない. これは, 2 つの成分 $X_{\mathrm{A}}[i,k]$, $X_{\mathrm{B}}[i,k]$ が, それぞれ複素正弦波であるため, 位相差によって加算後のエネルギが変化するから

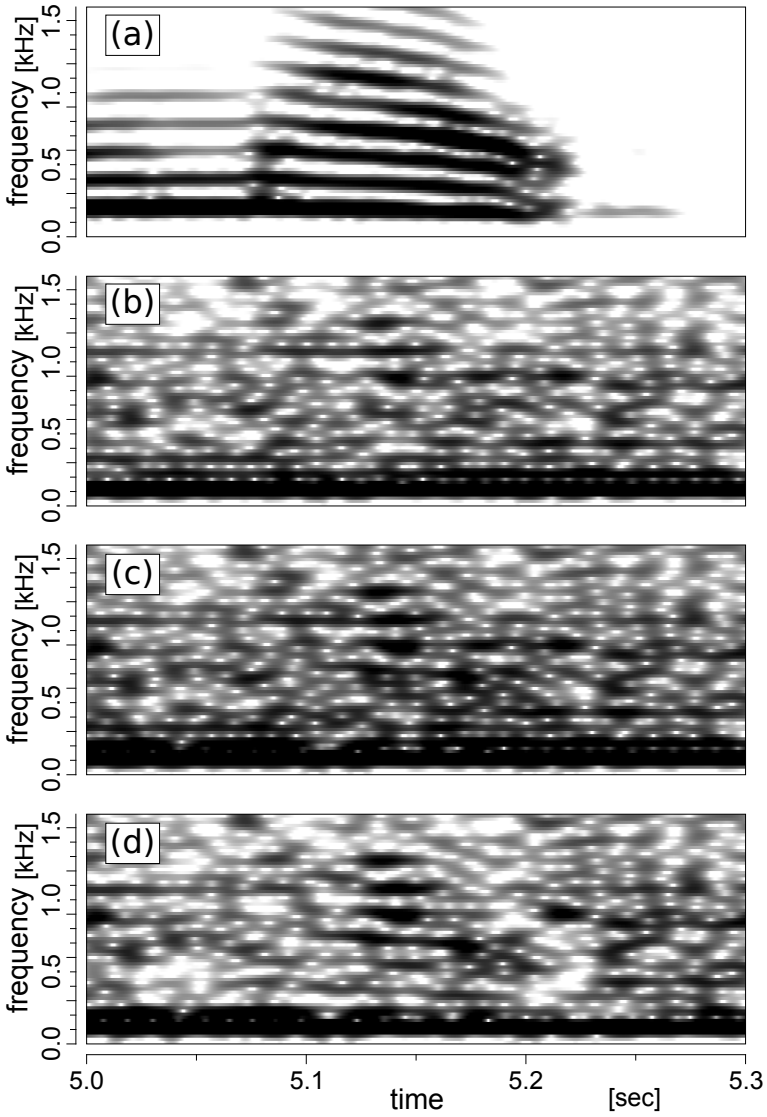


図 4.8. 各サンプルの時間周波数平面　(a) 音声 (b)BGM (c) 単純加算 (d) 局所抑制法 (手法 2)

である.

さて, 入力信号の内容や種類を判別するために重要なのは, スペクトルのピークの時系列である. このピークが, もう一方の入力信号の成分との位相差によっては, 最悪無音になることがある.

ミキシング時に入力信号の重要な成分が, 位相の兼ね合いでエネルギが削られるのを防ぐ方法として, 局所同期法を検討する.

4.1.2.1 物理的に無理のない位相の関係を保持する手法

局所同期法は, 従来法として高橋 [57, 58] による手法がある.

この手法では, 優先度が低い音信号の時間周波数ビン $[i, k]$ の位相を, 局所的に優先度

が高い音信号の時間周波数ビン $[i,k]$ の位相に同期させる．このとき, 位相を操作することによって, 時間周波数平面上の近接領域との位相の関係が崩れてしまう．大幅な崩れは聴取時に悪い印象を与える．音質のよい処理を行うためには, 物理的に無理のない位相の関係を保持する必要がある．

そこで, 各周波数ビン $[i,k]$ での位相操作の歪みを時間周波数平面上の近接領域 $[i,k-1]$, $[i,k+1]$, $[i-1,k]$, $[i+1,k]$ に滲ませる．このとき, 優先度の高い入力信号の有音なスペクトル上のピークの成分の領域では, 滲みが少なくなるように重みをつけて平滑化する．位相調整量が滲みによって小さくなることを防ぐことで, 位相同期の効果を維持する狙いである．平滑化を数回～数百回のイテレーション演算として繰り返し, 最終的に近接領域での位相の関係が釣り合う値に収束させる．

4.1.2.2 手法 3　音量差に着目した手法

入力信号の時間周波数平面上の音量関係によって, 位相同期処理を実施するかを決定する手法を, 手法 3 として検討する．

手法 3 は, 従来局所同期法でのピーク判定, 有音判定, 近接領域とのイテレーション演算を行わない, 単純な処理とする．位相操作の歪みを滲ませる処理を省略したため, 位相調整量の時間周波数平面上での滑らかさに乏しい．ミュージカルノイズの発生に伴う音質劣化をどう抑えるかが焦点となる．

そこで, 手法 3 は入力信号の時間周波数平面上の音量関係に対して, 位相操作の効果が大きい時間周波数成分にのみ処理を行うことを狙う．このような成分は, 音声や BGM のピーク成分に相当すると予想される．ピーク成分の時間周波数平面上での周縁の成分は, ピーク成分が充分に強調されることで, マスキング効果により目立たなくなることが期待できる．ピーク成分の時間周波数平面上での周縁での位相操作の歪みを, このマスキングの影に押し込んでしまうことを狙った．

処理について述べる． 検討する局所同期法を, 局所抑制法の変種として定義する．局所抑制法の非線形関数 $F(L_{\mathrm{A}}[i,k], L_{\mathrm{B}}[i,k])$ の出力を, ゲイン係数 $W_{\mathrm{B}}[i,k]$ に換えて, $X_{\mathrm{B}}[i,k]$ への同期係数 $M_{\mathrm{B}}[i,k]$ とし, 式 (4.6) で算出するものとする．ここで, $\phi_{\mathrm{A}}[i,k]$, $\phi_{\mathrm{B}}[i,k]$ は, それぞれ $X_{\mathrm{A}}[i,k]$, $X_{\mathrm{B}}[i,k]$ の位相である．

$$M_{\mathrm{B}}[i,k] = \begin{cases} \exp j(\phi_{\mathrm{A}}[i,k] - \phi_{\mathrm{B}}[i,k]) & l_{\mathrm{a}} < L_{\mathrm{A}}[i,k] < l_{\mathrm{A}}, l_{\mathrm{b}} < L_{\mathrm{B}}[i,k] < l_{\mathrm{B}} \\ 1 & \text{otherwise} \end{cases} \tag{4.6}$$

本手法では処理対象となった優先度が低い音信号の成分の位相を, 完全に優先度が高い音信号の成分の位相に置き換える．出力 $X_{\mathrm{S}}[i,k]$ は, $X_{\mathrm{S}}[i,k] = X_{\mathrm{A}}[i,k] + M_{\mathrm{B}}[i,k]X_{\mathrm{B}}[i,k]$ で決定する．

本手法で良好な処理結果が得られるパラメータの特徴を検討した．パラメータ調整は, 局所抑制法での周波数帯域の設定 (4.1.1 節, 表 4.3) と揃えた．

実データによる簡単な聴取実験による検討の結果, 以下の知見を得た．第一として, 音声の調波構造のピークの部分を同期させることで, 違和感の少ないミキシングが可能

である. 第二として, 音声の調波構造のピークの周縁部を同期させた場合, ミュージカルノイズが顕著となる.

4.1.3　音量関係に着目した局所抑制法と局所同期法の連携

本節では, 4.1.1 節で検討した局所抑制法 (手法 2) と 4.1.2 節で検討した局所同期法 (手法 3) を連携した処理を手法 4 とし, 検討する.

局所抑制法 (手法 2) と局所同期法 (手法 3) は, それぞれパラメータの数が 24 ある. これらのパラメータが取りうる値の組み合わせは膨大であり, 調整するのは困難である. そこで, 局所抑制法と局所同期法を独立に調整し, パラメータ組を連携する方法を検討する.

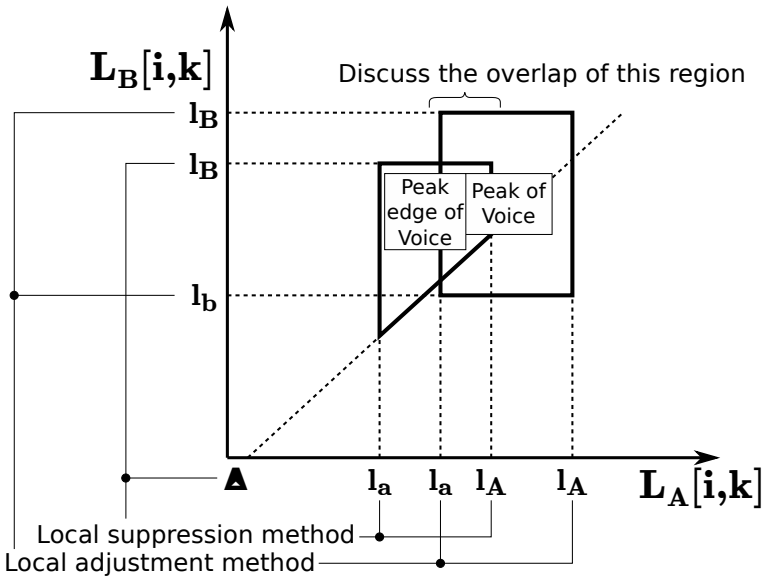


図 4.9. 局所抑制法 (Local suppression method) と局所同期法 (Local adjustment method) の良好な処理結果が得られるパラメータの模式図

ここで, 局所抑制法と局所同期法の良好な処理結果が得られるパラメータは, 図 4.9 に示すように, 前者では音声の調波構造のピークの周縁部とし, 後者では音声の調波構造のピークとする. 従って, 同一周波数帯域での局所抑制法と局所同期法のパラメータの値域の関係として現れるのは, 以下の 3 パターンである.

パターン 1 局所抑制法の $l_{\mathrm{A}} >$ 局所同期法の l_{a}

パターン 2 局所抑制法の $l_{\mathrm{A}} =$ 局所同期法の l_{a}

パターン 3 局所抑制法の $l_{\mathrm{A}} <$ 局所同期法の l_{a}

複数の音信号のセットについて, 実際にパラメータ調整を行った. 音声には SRV-DB[55] の「発話のプロフェッショナルによる編集手帳（読売新聞）の読み上げ」のデー

タを, 音楽には「RWC 研究用音楽データベース：ポピュラー音楽 [56]」のデータを用いた. その結果, 局所抑制法と局所同期法のパラメータの値域の関係は, パターン 1 と 3 とになった. 両パターンについて, 組み合わせ方を検討する.

まず, パターン 1 : 局所抑制法の $l_{\mathrm{A}} >$ 局所同期法の l_{a} である. このとき, パラメータを連携させる方法として, 次の 4 つの方法を試した.

A:振幅処理優先 局所同期法の l_{a} を, 局所抑制法の l_{A} に置き換える.

B:振幅位相供用 パラメータ変更なし. 値域が被る成分では, 両方の処理を行う.

C:位相処理優先 局所抑制法の l_{A} を, 局所同期法の l_{a} に置き換える.

D:振幅位相中間 局所抑制法の l_{A} と局所同期法の l_{a} を, それぞれの値の間の値に置き換える. 実験では, 簡単に中点の値とした.

処理を行った結果, C:位相処理優先が, 最も良好な結果となった. 位相処理を減らす, A:振幅処理優先と D:振幅位相中間では, 音声の埋もれ感が増強し, 処理結果に物足りなさを感じた. 一方, 振幅処理を増やす, B:振幅位相供用では, 音質劣化が耳についた.

次に, パターン 3 : 局所抑制法の $l_{\mathrm{A}} <$ 局所同期法の l_{a} である. このとき, パラメータを連携させる方法として, 次の 4 つの方法を試した.

A:振幅処理優先 局所抑制法の l_{A} を, 局所同期法の l_{a} に置き換える.

B:振幅位相供用 パラメータ変更なし. 値域が被らない成分では, 処理を行わない.

C:位相処理優先 局所抑制法の l_{A} を, 局所同期法の l_{a} に置き換える.

D:振幅位相中間 局所抑制法の l_{A} と局所同期法の l_{a} を, それぞれの値の間の値に置き換える. 実験では, 簡単に中点の値とした.

処理を行った結果, B:振幅位相供用が, 最も良好な結果となった. 位相処理を増やす, C:位相処理優先と D:振幅位相中間では, ボコッというノイズが増えた. 振幅処理を増やす, A:振幅処理優先では, 音質劣化が目立つ結果となった.

本手法についての結論を述べる. 振幅処理と位相処理を連携させる上で, 次の指針が得られた. それぞれ独立して調整したパラメータを, 少なくとも処理を増やす側に変更するとノイズが発生する. 一方の処理を増やしたことによって発生したノイズによる劣化した印象を, もう一方の処理の効果によって誤魔化すことは難しい.

4.2　調波構造に着目した手法

本節では, 調波構造に着目した手法を検討する. 調波構造のスペクトル遷移パターンを判定することで, パラメータ数の削減を狙う. また, 判定に際し, 聴覚心理モデルを導入する.

この聴覚心理モデルとしては, 音声符号化方式の一つである MPEG-1/Audio Layer1/2 で用いられている Psychoacoustic medel 1 を利用する [59, 60]. このモデルは, テレビ放送や Web 上のコンテンツでの使用実績がある. また, 同じ音声符号化方式である MP3(MPEG-1/Audio Layer3) の聴覚心理モデルと比較したとき, 遅延が少なく構造も簡潔である. リアルタイムでの動作を実装の目標とする本研究では, より簡易な MPEG-1/Audio Layer1/2 の聴覚心理モデルが適している.

MPEG-1/Audio Psychoacoustic model 1 の概略を, A 章に記した.

4.2.1　調波構造に着目した局所抑制法

本節では, 4.1 節で述べた局所抑制法 (手法 2) のゲイン係数生成を, 聴覚心理モデルで算出する特徴量によって行う 2 手法を検討する. 第一の手法が, 純音判定結果を利用する手法である. これを手法 5 とし, 4.2.1.2 節で述べる. 第二の手法が, 聴覚上の各成分の音量を利用する手法である. これを手法 6 とし, 4.2.1.3 節で述べる.

ここで, 純音とは聴覚心理モデルにおける以下の意味の成分を指す. 時間領域での幅を時間フレーム i とし, 周波数領域での幅を周波数ビン k を中心とする臨界帯域とする時間周波数領域内で特定の周波数にのみエネルギが集中している成分である.

4.2.1.1　局所抑制法の非線形関数への聴覚心理モデルの導入

4.1 節で検討した局所抑制法の非線形関数は, 各入力信号をエネルギで 3 つの成分に分けて捉えることを目指して構成されている. すなわち, 純音, 非純音, 非可聴音の 3 つの成分である. これらの成分分解を, 従来は時間方向に固定の閾値で行っていた. しかし, 入力信号は時変であるため, 非線形関数のパラメータも時変であることが望ましい.

そこで, 聴覚心理モデルを導入する. 聴覚心理モデルは, 聴覚上の特徴量を算出する. その値を用いることで, その時区間ごとに非線形関数のパラメータを適した値にできると考えた.

聴覚心理モデルを導入するにあたり, 従来の非線形関数の特性 (図 4.6) を, 表 4.5 のように解釈する. 図 4.6 での抑制を行う領域である実線枠は, 表 4.5 の実線枠内 (すなわち, 音声, BGM ともに非純音) で, BGM のエネルギが音声のエネルギに勝っている成分の集合となる.

4.2.1.2　手法 5 : 純音判定結果の利用

手法 2 の非線形関数では, 閾値 $l_{\mathrm{a}}, l_{\mathrm{A}}$ により, 入力信号を純音, 非純音, 非可聴音の 3 つの成分として分けて捉えることを狙っていた. この判別を, 聴覚心理モデルの判定に置き換える. これにより, 入力信号の入力レベルや周波数特性に, 自動的に適合できると思われる.

手法 2 での抑制量 $W_{\mathrm{B}}[i,k]$ は, 次のように書ける. フレーム i, 周波数ビン k において, 音声, BGM がともに純音ではなく, かつ, 音声, BGM の音量 $L_{\mathrm{A}}[i,k]$, $L_{\mathrm{B}}[i,k]$ が, 最小

表 4.5. 局所抑制法の非線形関数の特性 “○”: 処理を行う “×”: 処理を行わない

BGM	純音	×	×	×
	非純音	×	○	×
	非可聴音	×	×	×
		非可聴音	非純音	純音
		音声		

可聴値 ATH[k] よりも大きいとき, $W_{\mathrm{B}}[i,k] = L_{\mathrm{A}}[i,k] - L_{\mathrm{B}}[i,k]$ とし, それ以外のとき, $W_{\mathrm{B}}[i,k] = 0$ とする.

4.2.1.3 手法 6 : 聴覚上の各成分の音量の利用

非線形関数で音量の比較をするにあたり, 聴覚上の音量を用いることで, より聴感上で適した抑制が行えることが予想できる. そこで, 手法 5 での $L_{\mathrm{A}}[i,k]$, $L_{\mathrm{B}}[i,k]$ を, 入力信号それぞれの聴覚上の音量 $S_{\mathrm{A}}[i,k]$, $S_{\mathrm{B}}[i,k]$ で置き換える.

手法 6 での抑制量 $W_{\mathrm{B}}[i,k]$ は, 次のように書ける. フレーム i, 周波数ビン k において, 音声, BGM がともに純音ではなく, かつ, 音声, BGM がともに非純音のとき, $W_{\mathrm{B}}[i,k] = S_{\mathrm{A}}[i,k] - S_{\mathrm{B}}[i,k]$ とし, それ以外のとき, $W_{\mathrm{B}}[i,k] = 0$ とする.

$S_{\mathrm{A}}[i,k]$, $S_{\mathrm{B}}[i,k]$ は, 臨界帯域に合わせた周波数成分の間引き後の値であるため, 高い周波数になるにつれて, 広い周波数帯域で同一の値として扱われる. このため, 手法 6 は手法 5 よりも聴感特性をよりよく捉えていることが期待できる反面, 時間周波数平面上での緻密さを失っている.

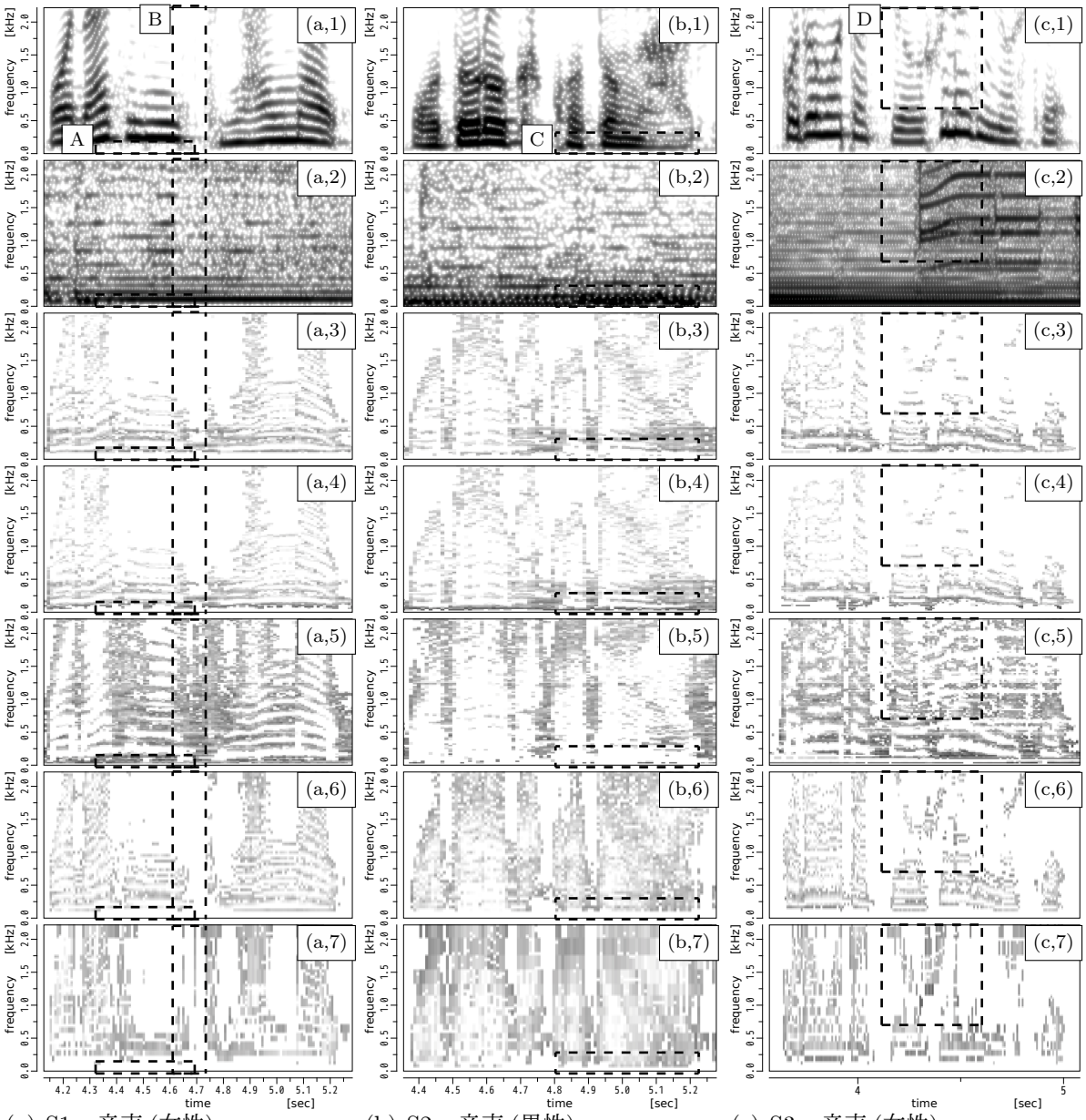
4.2.1.4 評価実験

評価実験として, 手法 5(4.2.1.2 節), 手法 6(4.2.1.3 節) で聴覚心理モデルに基づいて生成したゲイン係数と, 手法 2(4.1 節) の手動調整により生成したゲイン係数が, どれだけ類似しているかを評価する.

手法 2 と手法 5,6 それぞれの実験諸元は, 表 4.6 とした. 手法 5,6 は, MPEG-1 Audio Layer2 の仕様で実装した. 予備実験において Layer2 は, Layer1 の仕様での処理結果よりも優れていたためである. Layer1 では, 特に基本周波数の変動が大きい時区間で, 純音が適切に判定されず, 抑制がなされなかった.

実験で用いたサンプルの音データは, 音声と BGM がそれぞれ別音源の 3 組 (S1~3) とした. 音声には SRV-DB[55] の「発話のプロフェッショナルによる編集手帳（読売新聞）の読み上げ」のデータを, 音楽には「RWC 研究用音楽データベース：ポピュラー音楽 [56]」のデータを用いた.

ゲイン係数を観察する. 入力信号の時間周波数平面画像と, 各手法により生成されたゲイン係数を, 図 4.10 に示す.



(a) S1：音声 (女性),
BGM(J POP)

(b) S2：音声 (男性),
BGM(シティーロック)

(c) S3：音声 (女性),
BGM(ポップス)

図 4.10. 入力信号の時間周波数平面 (1:音声, 2:BGM) とゲイン係数 (3∼5:手法 2, 6:手法 5, 7:手法 6)

表 4.6. スマートミキサーの実験諸元

	手法 2	手法 5,6
サンプリング周波数	44.1 kHz	44.1 kHz
量子化ビット数	16 bit	16 bit
FFT 点数	4096 点	1024 点
解析時窓関数	ハニング窓形 2048 点	ハニング窓形 1024 点
合成時窓関数	ハニング窓形 1024 点	ハニング窓形 768 点
フレームシフト	512 点	384 点

図 4.10 は, 7 行 3 列の図であり, 列方向では同じ入力信号組での処理結果と時間周波数平面を配置し, 行方向では同じ手法での処理結果か時間周波数平面を配置している. 各行は, 1∼2 行：時間周波数平面 (1：音声, 2：BGM), 3∼5 行：手法 2 について各音素材の組について調整されたパラメータによるゲイン係数 (3：S1, 4：S2, 5：S3), 6∼7 行：聴覚心理モデルに基づいたゲインマスク (6：純音判定結果のみ利用 (手法 5), 7：聴覚上の音量を利用 (手法 6)) を配置した. 各列は, a 列：S1, b 列：S2, c 列：S3 での処理結果と時間周波数平面を配置した. 手法 2 の処理結果で, 入力信号組 (S1∼3) それぞれについて最適なパラメータでの処理結果は, S1 が (a,3), S2 が (b,4), S3 が (c,5) である. 対応するこれらのパラメータを, P1∼3 とする.

図 4.10 の破線枠を例にとり, 考察を行う. それぞれの時間周波数平面画像では, パワーまたは抑制量が大きいほど黒い. ゲイン係数 (a∼c,3∼7) は, 共通の閾値 −92 dB で正規化した.

[枠 A]　この枠内は, (a,1) の基本周波数以下の領域である. (a,2) では, この領域に成分が多く分布しており, 過剰な抑制では音質の劣化が知覚されてしまう. (a,3) では, 純音の分布に対して均一に抑制が行われている. (a,4) と (a,5) では, 基本周波数成分より下の成分を, 過剰に抑制してしまっている. (a,6) と (a,7) では, 全く判別できていない.

[枠 B]　この枠内は, (a,1) の無音区間である. この領域では抑制を行うべきではない. 音声と BGM のパワー差が大きいため抑制量が大きくなり, BGM のパワーが大きく変動してしまうためである. (a,3∼4) では, 0.5 kHz 以下の成分を抑制し続けてしまっている. (a,5) では, 純音の分布は捉えられているが, 過剰に抑制されてしまっている. (a,6) では, 0.5 kHz 以下でも適切に抑制している. (a,7) では, 0.5 kHz 以下では適切に抑制しているが, 2 度インパルス状に抑制している領域ができてしまっている.

[枠 C]　この枠内は, (b,1) の基本周波数成分が存在する. (b,4) では, 純音の分布が適切に判別できている. (b,3) と (b,6) では, 純音のパワーが弱い部分の検出が不十分であり, 音声の埋もれを解消できない. (b,5) では, 過剰に抑制してしまっている. (b,7) では, 全く判別できていない.

[枠 D]　この枠内は, (c,1) のフォルマントが存在する. (c,5) では, フォルマントのみならず, 調波構造がはっきりするまで抑制してしまっている. 他の処理結果に比べて抑制量が大きいため, BGM の劣化は激しい. しかし, この入力信号組は音量差が極端に大

表 4.7. 抑制後の BGM の時間周波数平面間の dB 値の差の絶対値の平均

手法＼信号組	S1 (dB)	S2 (dB)	S3 (dB)
手法 2 パラメータ P1	-	1.19	0.76
手法 2 パラメータ P2	0.73	-	0.82
手法 2 パラメータ P3	1.64	1.48	-
手法 5	0.84	1.93	0.85
手法 6	2.03	3.21	1.30

きい. この為, 音声を聞き取るために, 抑制を強くする必要があったのである. (c,3) と (c,4) では, 充分な抑制が行われていない. (c,6) と (c,7) では, 抑制量は少ないが, フォルマントは捉えられており, 音声の埋もれの解消に効果が期待できる.

手法別に観察する. 手法 2 の固定のパラメータでは, 調整対象の入力信号組以外では, 成分の分布を充分に捉えきれていない. 対して手法 5,6 では, 各ビンでの抑制量が一様で, 手調整に比べて大きくなる傾向になる. とくに, 高い周波数帯域になるほど, 多く抑制している. 2 kHz 以下に限れば, どの入力信号組に対しても, 適切に成分の分布を捉えられている.

続いて, 評価関数を用いて評価する. ゲイン係数の評価として, 抑制後の BGM の時間周波数平面間の, 各時間周波数ビンの dB 値の差の絶対値の平均を, 全周波数成分を対象として算出した. 手法 2 で信号組に対して手調整したパラメータでの処理結果と, それ以外の手法での処理結果を評価する. 評価する時間周波数平面は, 時間領域信号から同一の諸元 (手法 5,6 と同じ設定) で再び生成した. これは, 手法 2 と手法 5,6 とで, 実験諸元が異なるため, ゲイン係数同士を単純に比較できないことへの対処である.

算出結果を, 表 4.7 に示す.

表 4.7 と図 4.10 を比較して考察する. 手法 5,6 は, ゲイン係数で着目をした 2 kHz より低い帯域では, どの信号組に対しても平均的に特徴を捉えられている. しかし, 全時間周波数ビンを対象とした差の絶対値の平均では, 手法 2 に及ばなかった.

この理由として, 高い周波数の領域で, より多く抑制していることが上げられる. 聴覚心理モデルでは, 高い周波数成分になるほど間引きが広くなり, 処理がおおまかになる. このため, わずかな非純音によって広い領域の抑制が行われ, 総量としての抑制量が大きくなっている.

改良案として, より周波数分解能を細かくすることが上げられる. これにより, 高い周波数帯域で広く一様に抑制することを防ぐことができる.

さらに, 純音の分布とマスキング量の情報を用いることで, 最小限の抑制量のゲイン係数が生成できると思われる.

また, 今回の手法 5,6 では, 全帯域をみたときのスペクトルのバランスの変化, 前後の

時間との関係などの, より広い時間周波数平面での考慮をしていない. この考慮を行うことで, より自然な処理結果が得られることが期待できる.

4.2.1.5 局所抑制法の改良

局所抑制法 (手法 5,6) の改良を試みる.

第一の改良として, 純音を基準にした抑制条件を検討する. 子音部の過度な抑制を抑制する改良を試みる. 可聴な非純音成分全体で抑制を行なっていた. しかし抑制されるべきは, 純音成分への影響がある成分のみである.

そこで, 純音判定を基準にした抑制条件の追加を検討した. 検討した抑制条件は, 以下の 3 つである. 第一に, 純音判定の比較に用いる周波数帯域 (パラメータなし). 第二に, 純音から上下任意 bark の帯域 (パラメータ 2 つ). 第三に, 純音に隣接する任意の数の上下周波数ビン (パラメータ 2 つ).

この 3 つについて処理を行った結果, 第一の純音判定の比較に用いる周波数帯域 (パラメータなし) の制限を行ったとき, 50 Hz から 5 kHz 程度まで, バランスよく抑制することができた.

残りの 2 つ (第二の純音から上下任意 bark の帯域, 第三の純音に隣接する任意の数の上下周波数ビン) については, パラメータ調整が必要にも拘らず, どのパラメータ組でも, 低音域 (1 kHz より低い周波数帯域) と高音域 (2 kHz より高い周波数帯域) とで, 抑制が行われる領域のバランスをとることができなかった.

第二の改良として, 抑制量減少量制限を検討する. 手法 5,6 で知覚される音質劣化感の一つに, 過剰な音量変動感がある. これは, 局所的な抑制処理が原因である. そこで, 時間方向に滑らかに抑制量が変化するような工夫を施すことにした.

時間方向に滑らかにするには, 前後フレームでの抑制量と平均をとるような処理を行えばよい. このとき, 未来方向に隣接するフレームを考慮に入れると, リアルタイムでの処理時にデータの入力待ち時間が増える. ゆえに, 過去方向との関係性だけで処理を完結させる.

未来方向に隣接フレームを考慮に入れないことの弊害は, 近い未来に優先度が高い音信号の重要な成分が入力される時, 経時マスキングの考慮ができないことである.

具体的な抑制量減少量への制限の付け方として, 前フレームからの抑制量減少量が制限より多い時, 制限の減少量とすることを考えた. 急激な抑制量の増加は許すが, 急激な抑制量の減少は許さないということである. この制限により滑らかな音量変化とし, ノイズ抑制を狙う.

周波数ビン k について, 過去方向の隣接フレーム $i-1$ での抑制量 $W_B[i-1,k]$ と, 現在フレーム i での抑制量 $W_{\mathrm{B}}[i,k]$ の差分 $\Delta W_{\mathrm{B}}[i,k]$ を算出. $\Delta W_{\mathrm{B}}[i,k] = W_{\mathrm{B}}[i,k] - W_B[i-1,k]$ とする.

抑制量減少量制限値 $\Delta W_{\mathrm{B}}[i,k]^L$ を設定し, $\Delta W_{\mathrm{B}}[i,k]^L < \Delta W_{\mathrm{B}}[i,k]$ のとき, 抑制量 $W_{\mathrm{B}}[i,k] = \Delta W_{\mathrm{B}}[i,k]^L$ とする. ただし, $X_{\mathrm{A}}[i,k]$ が純音であるとき, 抑制量 $W_{\mathrm{B}}[i,k] = 0$ とする.

表 4.8. 局所同期法の非線形関数の特性　“○”: 処理を行う “×”: 処理を行わない

B G M	純音	×	×	○
	非純音	×	×	○
	非可聴音	×	×	×
		非可聴音	非純音	純音
		音声		

この処理により, 抑制による BGM の劣化感を好みに応じて調整できるようになった. 従来ミキシング処理では, コンプレッサの Release パラメータに相当する.

4.2.1.6 まとめ

本節では, 局所抑制法 (手法 2) の改良法を検討した. 改良法である手法 5, 6 では, 局所抑制法のゲイン係数を MPEG-1 Audio Layer1/2 の聴覚心理モデルの純音判定に基づいて生成する方法を示した. 評価として, 従来法と手法それぞれで生成したゲイン係数を観察による比較と, ゲイン係数による抑制後の BGM の時間周波数平面間の距離の算出を行った.

評価結果から, 聴覚心理モデルの導入により, 入力信号の分布を捉えたゲイン係数が生成できることが確認できた. しかし, 抑制量全体としては過剰であった. 特に高い周波数帯域で抑制量が多くなる傾向が確認された.

4.2.2　調波構造に着目した局所同期法

本節では, 4.1 章で述べた局所同期法 (手法 3) の同期係数の生成を, 聴覚心理モデルで算出する特徴量によって行う方法を検討する.

4.2.2.1 局所同期法の非線形関数への聴覚心理モデルの導入

4.1 章で検討した局所同期法 (手法 3) の非線形関数も, 局所抑制法 (手法 2) の非線形関数と同様に, 各入力信号を純音, 非純音, 非可聴音の 3 つの成分に分けて捉えることができる. そこで, 局所同期法についても 4.2.1 節で述べた局所抑制法と同様の手順で, MPEG-1 Audio Layer2 の聴覚心理モデルを導入する.

聴覚心理モデルを導入するにあたり, 従来の非線形関数の特性を, 表 4.8 のように解釈する. 入力信号のパワーの閾値 $l_{\mathrm{a}}, l_{\mathrm{A}}, l_{\mathrm{b}}, l_{\mathrm{B}}$ で仕切られる領域を, 純音, 非純音, 非可聴音の領域であると解釈する. この読み替えにより, 位相処理を行う領域は表 4.8 の実線枠内 (すなわち, 音声が純音, BGM が純音か非純音) となる.

表 4.9. 局所抑制法と局所同期法の良好な組み合わせ法
“局所抑制法”: 局所抑制法を行う　“局所同期法”: 局所同期法を行う　“×”: 処理を行わない

BGM	純音	×	×	局所同期法
	非純音	×	局所抑制法	局所同期法
	非可聴音	×	×	×
		非可聴音	非純音	純音
		音声		

4.2.2.2 手法 7 : 純音判定結果の利用

局所同期法の非線形関数では, 閾値 $l_{\mathrm{a}}, l_{\mathrm{A}}, l_{\mathrm{b}}, l_{\mathrm{B}}$ により, 入力信号を純音, 非純音, 非可聴音の 3 つの成分として, 分けて捉えることを狙っていた. この判別を, 聴覚心理モデルの判定に置き換える. これにより, 入力信号の入力レベルや周波数特性に自動的に適合できることが期待できる. この手法を手法 7 として検討する.

本手法では, 同期係数 $M_{\mathrm{B}}[i,k]$ を, 次のように算出する. フレーム i, 周波数ビン k とその上下 1 ビンのうち 1 ビン以上の音声が純音であり, かつ, 音声, BGM のそれぞれの音量 $L_{\mathrm{A}}[i,k]$, $L_{\mathrm{B}}[i,k]$ が, 最小可聴値 $\mathrm{ATH}[k]$ よりも大きいとき, $M_{\mathrm{B}}[i,k] = \exp j(\phi_{\mathrm{A}}[i,k] - \phi_{\mathrm{B}}[i,k])$ とし, それ以外のとき, $M_{\mathrm{B}}[i,k] = 1$ とする.

音声の純音として上下 1 ビンのズレを許容しているのは, 聴覚心理モデルが純音の聴覚上の音量として, 上下 1 ビンのパワーも含めた値としていることに従った.

4.2.3 調波構造に着目した局所抑制法と局所同期法の連携

手法 8 として, 聴覚心理モデルを導入した局所抑制法 (手法 5) と局所同期法 (手法 7) とを連携した手法を検討する.

局所抑制法の処理条件 (表 4.5) と, 局所同期法の処理条件 (表 4.8) を組み合わせると, 表 4.9 となる.

4.2.3.1 処理結果

手法 8 での処理結果では, 位相処理のみを行った処理結果に比べて低音が膨張したように聴こえる傾向があった. 膨張感のある時区間では, 発音が濁ってしまい音声の発話内容の把握に差し障りがあると感じた. 埋もれの解消については, 振幅処理, 位相処理をそれぞれ単体で行った処理結果に比べて, いずれのサンプルでも大幅に改善されて聴こえた.

低音が膨張したように聴こえる処理となった理由として, 第一に有音な純音成分が低音に集中していること, 第二に各入力信号で低音のパワーの割合が高いことが考えられる.

表 4.10. スマートミキサーの実験諸元

サンプリング周波数	44.1 kHz
量子化ビット数	16 bit
FFT 点数	1024 点
解析時窓関数	ハニング窓形 1024 点
合成時窓関数	ハニング窓形 768 点
フレームシフト	384 点

各入力信号のパワーの大きい帯域のパワーが, 位相同期法によって確実に加算された時, 高い周波数のパワーとのパワー比が大きくなる.

処理結果を改良する方法として, 以下の 3 つが考えられる. 第一に, 手法 8 で純音として捉える領域として上下周波数ビンとしているところを見直す. 第二に, より高い周波数帯域でも処理が行えるよう, 純音判定や有音判定を調整する. 第三に, 音の印象に関する特徴量 (例えば, スペクトルセントロイド) を保持するような機構をもたせる.

4.2.3.2 処理例と考察

本節で検討した手法での処理例を示し, 考察する. 特に, 手法 8 における局所抑制法に局所同期法を追加する効果に着目する.

検討するスマートミキサーの実験諸元は, 表 4.10 とした.

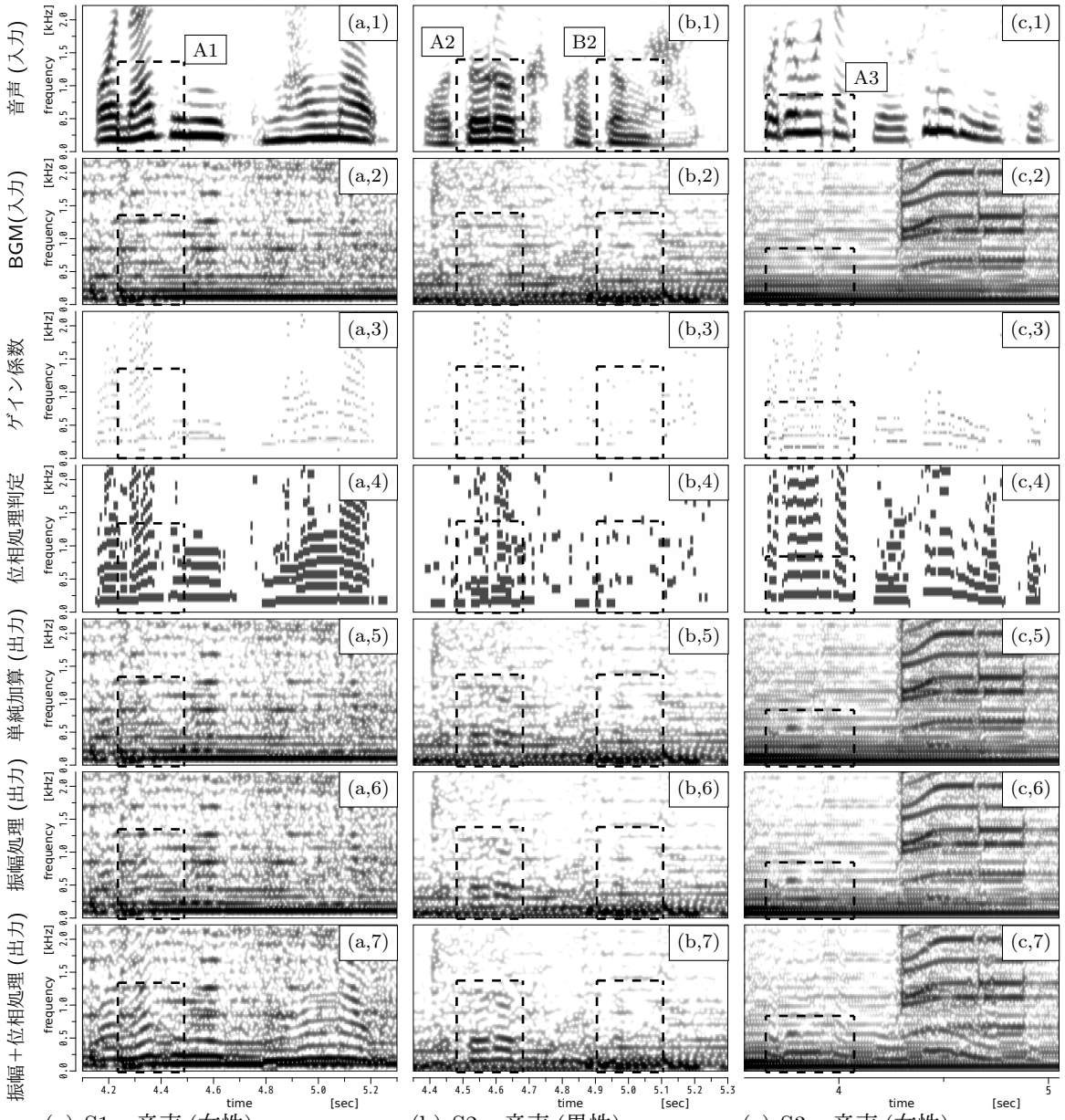
局所抑制法では, 音声の純音判定で考慮する近傍領域内の成分のみを, 処理対象とする条件を追加する改良を加えた処理とした.

局所同期法では, 処理を行う成分を, 純音と判定された時間周波数成分と, その上下周波数方向に隣接する成分とした.

時間周波数平面を観察する. 入出力信号の時間周波数平面画像と, 手法 8 によって生成された, ゲイン係数及び位相処理判定を, 図 4.11 に 7 行 3 列の図として示す. 列方向には, 同じ入力信号セット (列 a：S1, 列 b：S2, 列 c：S3) の画像を配置した. 行方向には, 同じ種類の画像を配置した. 上から, 入力信号の時間周波数平面 (行 1：音声, 行 2：BGM), 局所抑制法のゲイン係数 (行 3), 局所同期法の位相処理判定 (行 4), 出力信号の時間周波数平面 (行 5：単純加算, 行 6：振幅処理のみ, 行 7：振幅＋位相処理 (手法 8)) である. これらの時間周波数平面画像では, パワーまたは処理量が大きいほど黒い. 行 3 のゲイン係数は, 共通の閾値 −92 dB で正規化した. 行 4 の位相処理判定は, 白 (処理 on) と黒 (処理 off) の 2 値である.

図 4.11 の破線枠を例にとり, 考察を行う.

破線枠 A1∼3 の枠内では, 狙い通りの処理結果となっている. ゲイン係数 (行 3), 位相処理判定 (行 4) ともに, 調波構造の特徴を判別できている. 一方, 単純加算 (行 5) とゲイン係数の適用結果 (行 6) を比較した時, 後者では白い (パワーの弱い) 領域が増えている. この変化により, 音声 (行 1) のパワーの分布の濃淡の特徴が, 際立っている. さらに, 位相処理を施す (行 7) ことで, パワーの小さい純音成分まで際立った.



(a) S1： 音声 (女性), BGM(J POP)　(b) S2： 音声 (男性), BGM(シティーロック)　(c) S3： 音声 (女性), BGM(ポップス)

図 4.11. 入力信号の時間周波数平面 (1:音声, 2:BGM), ゲイン係数 (3), 位相処理判定 (4), 処理結果の時間周波数平面 (5:単純加算, 6:振幅処理のみ, 7:振幅処理と位相処理の連携 (手法 8))

一方, 破線枠 B2 の枠内では, 狙いから外れた処理結果となっている. 音声の時間周波数平面 (行 1) と位相処理判定 (行 4) を比較すると, 調波構造の特徴が捉えられていないことがわかる. 音声の時間周波数平面 (行 1) より, この領域の基本周波数は, 特に低い. 基本周波数が低いと, 周波数方向でのパワーの大小の間隔が狭い. このため, 純音判定に必要であった周波数帯域内でのパワー差が, 埋もれていると考えられる. 表 4.10 の諸元での FFT 点数 (1024 点) では, 周波数分解能が不足することがわかった.

4.3　音質に配慮した調波構造に着目した手法 (提案法 A)

本章では, 4.2 節での検討で得た知見を元に, 音質に配慮した調波構造に着目した手法を提案法 A として提案する. 4.2 節で検討した手法 8 は, 入力信号の調波構造のスペクトル遷移パターンの維持を目的とし, 時間周波数平面の各ビンについて振幅処理と位相同期処理を行う. 評価実験によって, 音声の埋もれなさは確保できる一方で音質劣化が過度であることがわかっている.

4.3.1　音質劣化低減の発想

混合後でも音声が埋もれずに聞き取れる条件として, 音声のスペクトル遷移パターンが, 混合後でも維持されることが重要であると考えた. 提案法 A では, スペクトル遷移パターンで重要な要素として調波構造に着目した. 音声の調波構造が, ミキシング後での時間周波数平面上でも調波構造となるように, 両信号へのゲインを計画する.

本手法では, 調波構造を捉える機構として, 音声圧縮の 1 手法である MPEG-1 Audio の聴覚心理モデルを用いる. MPEG-1 Audio では帯域ごとのビット割り当てを, 聴覚上で優位な純音が存在する帯域に優先的に多く配分する.

知見として, 音声と BGM の音量が同じであると音声が埋もれず聞き取れることが, 予備実験によってわかっている. そこで, 本手法では音声の重要な周波数帯域で, 音声と BGM が同じ音量にするのを目標値としている.

一方, 音声の明瞭さは, 調波構造の先鋭化で向上できる. 本手法では, 純音と非純音の音量差を強調することで, 調波構造を先鋭化する.

各周波数ビンのみでは, 定常な周波数の正弦波である. 近傍領域との兼ね合いによって, その周波数ビンの成分は近い周波数帯域の任意な信号を表現できる. ここで, 近傍領域の寄与率は窓関数の周波数特性で決まる. 本手法では, 考慮する近傍領域を MPEG-1 Audio の聴覚心理モデルに従って設定した.

4.3.2　提案法 A の処理

本手法のブロック図を, 図 4.12 に示す. 本手法は, 2 つのモノラル入力 (優先度の高い順に $x_{\mathrm{A}}[n]$, $x_{\mathrm{B}}[n]$) と, 1 つのモノラル出力 $y[n]$ を有する. ここで, n を時間サンプリング番号とする. 音声と BGM とを混合する場合には, 音声が優先度の高い入力信号であ

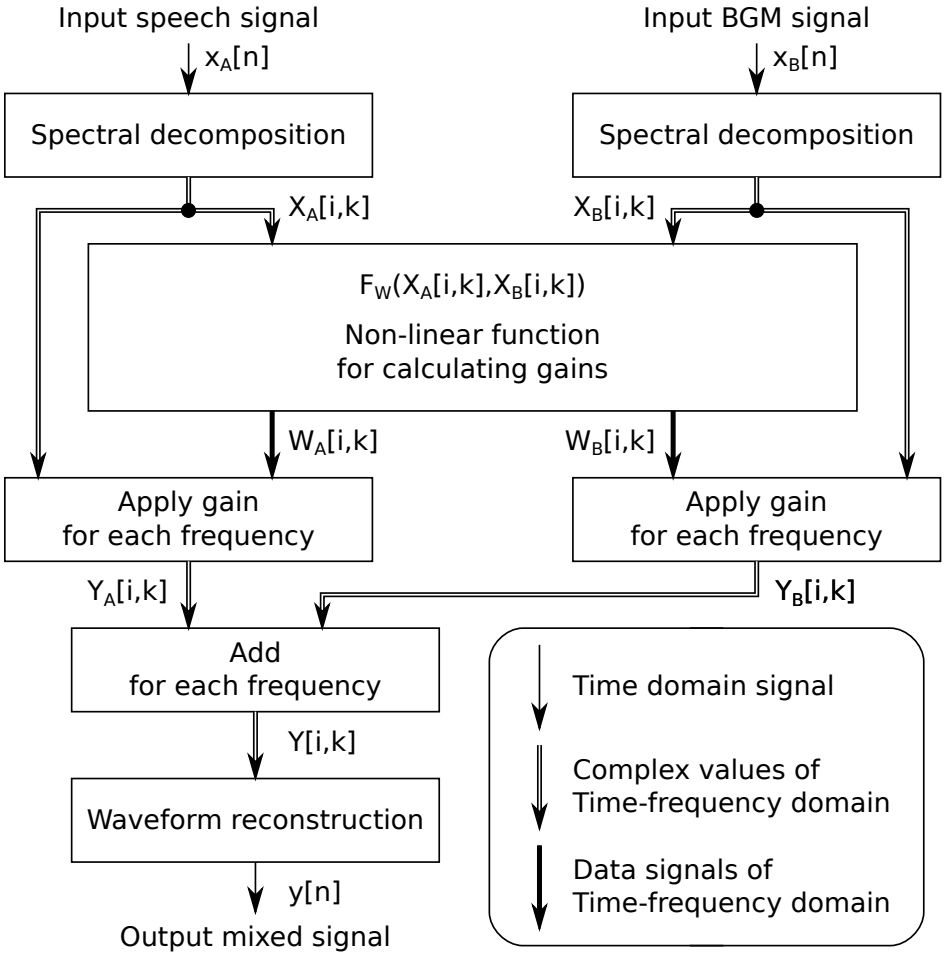


図 4.12. 提案法 A のブロック図

り, BGM が優先度の低い入力信号である. $X_{\mathrm{A}}[i,k]$, $X_{\mathrm{B}}[i,k]$, $Y[i,k]$ は, 各信号の時間周波数平面上の位置 $[i,k]$ での, 短時間フーリエ変換値である. ここで, i は時間フレーム番号, k は周波数ビン番号である. $X_{\mathrm{A}}[i,k]$ へのゲイン係数 $W_{\mathrm{A}}[i,k]$, $X_{\mathrm{B}}[i,k]$ へのゲイン係数 $W_{\mathrm{B}}[i,k]$ を算出する非線形関数を, $F_W(X_{\mathrm{A}}[i,k], X_{\mathrm{B}}[i,k])$ とする. 続いて, $Y_{\mathrm{A}}[i,k]$ と $Y_{\mathrm{B}}[i,k]$ を $X_{\mathrm{A}}[i,k]$ と $X_{\mathrm{B}}[i,k]$ に $W_{\mathrm{A}}[i,k]$ と $W_{\mathrm{B}}[i,k]$ をそれぞれ乗算することで生成し, 出力信号の時間周波数表現 $Y[i,k]$ を $Y[i,k] = Y_{\mathrm{A}}[i,k] + Y_{\mathrm{B}}[i,k]$ とする. 最後に, この $Y[i,k]$ への逆短時間フーリエ変換により, 出力信号 $y[n]$ を得る.

非線形関数 $F_W(X_{\mathrm{A}}[i,k], X_{\mathrm{B}}[i,k])$ について述べる. 音声を埋もれさせない目的を達成するために, 手法では音の知覚に焦点を当てた. 音声と BGM の知覚上の違いは, 時間周波数分布の時間変化パターンに関係している. 聴覚上重要な時間周波数成分を抽出するための方法の一つとして, MPEG-1 audio に含まれる聴覚心理モデル [59, 60] がある. 入力信号を聞くために重要な成分の音質を維持するという点で, 音声データの圧縮と手法は目的が共通している. 従って, 本手法は成分を強調, 抑制または不変とすべきかを判断するために, MPEG-1 Audio の聴覚心理モデルに含まれる純音判定と有音判

定を用いる. なお可聴判定は, 入力信号のレベルに追随するために拡張した.

純音判定 $C_{\mathrm{tm}}[X,i,k]$ (Boolean) は, 式 (4.7), (4.8) で与えられる. ここで, $L_{\mathrm{dB}}[X,i,k]$ は $X[i,k]$ の dB 値である.

$$C_{\mathrm{tm}}[X,i,k] = \begin{cases} 1 & \left(\begin{array}{c} L_{\mathrm{dB}}[X,i,k] > L_{\mathrm{dB}}[X,i,k'] \\ \wedge\, L_{\mathrm{dB}}[X,i,k] > L_{\mathrm{dB}}[X,i,k''] + 7, \\ \forall k' \in k + \{-1,1\}, \forall k'' \in k + \Delta_k \end{array}\right), \\ 0 & (\text{otherwise}) \end{cases} \tag{4.7}$$

$$\Delta_k = \begin{cases} -2, 2 & (2 < k < 63) \\ -3, -2, 2, 3 & (63 \le k < 127) \\ -6, \cdots, -2, 2, \cdots, 6 & (127 \le k < 255) \\ -12, \cdots, -2, 2, \cdots, 12 & (255 \le k \le 500) \end{cases}. \tag{4.8}$$

Δ_k は, 周波数帯域で値が異なる. ここに, 臨界帯域の周波数帯域での拡がりの違いが反映されている.

有音判定は, MPEG1 Audio Psychoacoustic model 1[59, 60, 61, 62] の可聴判定を元にした. 可聴判定は, 最小可聴値 (ATH, absolute threshold of hearing)[60] の音量値との比較で行われる. 最小可聴値は, 式 (4.9) と (4.10) で与えられる.

$$\begin{aligned} \mathrm{ATH}[f] = 3.64\,(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} \\ +10^{-3}\,(f/1000)^4 \quad (\text{dB SPL}), \end{aligned} \tag{4.9}$$

$$C_{\mathrm{ATH}}[X,i,k] = \begin{cases} 1 & \left(\begin{array}{c} L_{\mathrm{dB}}[X,i,k] + p_{\mathrm{ATH}} > \\ \mathrm{ATH}[k] + L^A_{\mathrm{dB}}[X,i] - p_{\mathrm{range}} \\ \wedge L_{\mathrm{dB}}[X,i,k] - p_{\mathrm{ATH}} > 0 \end{array}\right). \\ 0 & (\text{otherwise}) \end{cases} \tag{4.10}$$

ここで, $L^A[X,i]$ は $X[i,k]$ の全周波数ビンでの平均時間 p_{t}^A(本稿では 60 ms) での平均エネルギ, $L^A_{\mathrm{dB}}[X,i]$ は $L^A[X,i]$ の dB 値, p_{ATH}(本稿では -69) は入力信号のダイナミックレンジとリスニングレベルの差への調整パラメータ, p_{range}(本稿では 96) は考慮する入力信号のダイナミックレンジを調整するパラメータである. これらの変数の導入により, 効果の深さを入力信号の平均レベルの時間変動に対応させることを狙う.

次に, $X_{\mathrm{A}}[i,k]$, $X_{\mathrm{B}}[i,k]$ を, 純音判定と有音判定の組み合わせから, D_{tm}, D_{tn}, D_{nm}, D_{na} の 4 種類に分類する. D_{tm}, D_{tn}, D_{nm}, D_{na} は, その種類であるか否かの判定結果とし, Boolean 値で表すものとする. ここで添字は, 以下の用語の略語とする.

表 4.11. 心理音響モデルに基づいた成分の分類による, 4 つの処理を切り替えるための条件のテーブル

		Priority input : voice (X_{A})			
		D_{na}	D_{nm}	D_{tn}	D_{tm}
Nonpriority input	$\neg D_{\mathrm{na}}$		S_{nm}	S_{tn}	S_{tm}
: BGM (X_{B})	D_{na}	S_{na}			

D_{tm}　the tonal masker

D_{tn}　the neighbor of the tonal masker

D_{nm}　the noise masker excluding tonal masker and neighbor of tonal masker

D_{na}　not audible

4 種類の判定は, 式 (4.11)-(4.14) で得る.

$$D_{\mathrm{tm}}[X,i,k] = \sum_{k'''\in\{0,\pm1\}} \begin{pmatrix} C_{\mathrm{tm}}[X,i,k+k'''] \\ \wedge C_{\mathrm{ATH}}[X,i,k+k'''] \end{pmatrix}, \tag{4.11}$$

$$D_{\mathrm{tn}}[X,i,k] = \sum_{k'''\in\Delta_k} \begin{pmatrix} C_{\mathrm{tm}}[X,i,k+k'''] \\ \wedge C_{\mathrm{ATH}}[X,i,k+k'''] \end{pmatrix} \wedge \neg D_{\mathrm{tm}}[X,i,k], \tag{4.12}$$

$$D_{\mathrm{nm}}[X,i,k] = C_{\mathrm{ATH}}[X,i,k] \wedge \neg D_{\mathrm{tm}}[X,i,k] \wedge \neg D_{\mathrm{tn}}[X,i,k], \tag{4.13}$$

$$D_{\mathrm{na}}[X,i,k] = \neg C_{\mathrm{ATH}}[X,i,k] \wedge \neg D_{\mathrm{tm}}[X,i,k] \wedge \neg D_{\mathrm{tn}}[X,i,k]. \tag{4.14}$$

4 種類の成分はそれぞれ以下の意味合いを持つ. D_{tm} は, 解析窓関数のメインローブに相当する成分である. D_{tn} と D_{nm} は, 有音な非純音成分である. D_{tn} は純音を中心とする臨界帯域内の成分であり, D_{nm} は臨界帯域内に有音な純音が存在しない成分である. D_{tn} を設定することで, 純音を分離して知覚させるための臨界帯域内での非純音成分との音量差を操作できる. D_{na} は非可聴でかつ近傍領域内に有音な純音が存在しない領域である.

続いて, 入力信号間での 4 種類の成分判定 D_{tm}, D_{tn}, D_{nm}, D_{na} の組み合わせを算出する. ゲイン係数の算出は, 各入力信号の分類結果の組み合わせで異なる関数で行う. 表 4.11 に, 各入力信号の分類結果の組み合わせと関数の対応を示す. これらの組み合わせごとに, 関数を異なるものとすることで, 分類結果の間に音量差を作り出す. 分類結果での関係が, 音声の重要な構成要素がミキシング後に聞こえるかどうかを決定するからである.

$W_{\mathrm{A}}[i,k]$, $W_{\mathrm{B}}[i,k]$ は, 式 (4.16)-(4.18) で算出する. $F_{\mathrm{lim}}(g,l,u)$ はリミット関数であり,

表 4.12. g_{A}, g_{B} の設定

Situations	S_{tm}	S_{tn}	S_{nm}	S_{na}
g_{A}	g_{tm}	1	g_{nm}	$\frac{L^f_{\mathrm{A}}[i,k]}{\left(L^f_{\mathrm{A}}[i,k]+L^f_{\mathrm{B}}[i,k]\right)}$
g_{B}	0	0	g_{nm}	$\frac{L^f_{\mathrm{B}}[i,k]}{\left(L^f_{\mathrm{A}}[i,k]+L^f_{\mathrm{B}}[i,k]\right)}$

上限 u と下限 l により, g を u と l の間の値に制限する.

$$F_{\mathrm{lim}}(g,l,u)=\begin{cases}u & u<g\\ g & l\le g\le u\\ l & g<l\end{cases} \tag{4.15}$$

$F_{\mathrm{lim}}(g,l,u)$ の導入により, 急激な強調および抑制を制限する.

$$W_{\mathrm{A}}[i,k]=F_{\mathrm{lim}}\left(\sqrt{g_{\mathrm{A}}\cdot\frac{L^f[X_{\mathrm{A}},i,k]+L^f[X_{\mathrm{B}},i,k]}{L^f[X_{\mathrm{A}},i,k]}},\ 1,\ l_{\mathrm{af}}\right), \tag{4.16}$$

$$W_{\mathrm{B}}[i,k]=F_{\mathrm{lim}}\left(\sqrt{g_{\mathrm{B}}\cdot\frac{L^f[X_{\mathrm{A}},i,k]+L^f[X_{\mathrm{B}},i,k]}{L^f[X_{\mathrm{B}},i,k]}},\ 1/l_{\mathrm{af}},\ 1\right), \tag{4.17}$$

$$l_{\mathrm{af}}=F_{\mathrm{lim}}\left(\sqrt{L^A[X_{\mathrm{B}},i]/L^A[X_{\mathrm{A}},i]},\ 1,\ l_{\mathrm{pb}}\right). \tag{4.18}$$

ミュージカルノイズ低減を狙っての工夫として, 時間周波数平面上でなだらかなゲイン係数を生成する. 手法では, ゲイン係数算出元のパワーの時間周波数平面の時間方向, 周波数方向の双方で平滑化と, 算出後のゲイン係数を時間方向に平滑化の 2 策を併用する.

$L^f[X,i,k]$ は時間周波数表現の近傍周波数帯での平均化パワーであり, 帯域幅は予備実験より 1 オクターブとした. l_{af} は, スペクトル包絡に対する極端な強調を避けるための, ゲイン係数の制限パラメータである. l_{pb}(本稿では, 4.8) は, 総パワーの急激な変調を避けるための, ゲイン係数への制限パラメータである. g_{A}, g_{B} は, 混合後の各時間周波数成分において, いずれかの入力信号に類似するかのバランスを調整するゲイン係数である. g_{A}, g_{B} の設定を, 表 4.12 に示す. これらの設定で良い効果が得られることを, 予備実験により確認している. ここで, g_{tm}(本稿では 4), g_{nm}(本稿では 0.8) は, 対応する成分でのゲインである.

最後に $W_{\mathrm{A}}[i,k]$, $W_{\mathrm{B}}[i,k]$ は, 平均時間 p_{t}(本稿では 60 ms) で平滑化する. この平滑化により, ミュージカルノイズが改善される.

表 4.13. 入力信号組 3 セットの設定

Set		Contents	Standard deviation	Volume difference
1	Voice	PF00 (Female)	500	12 dB
	BGM	No.5	2000	
2	Voice	PM00 (Male)	1000	12 dB
	BGM	No.15	4000	
3	Voice	PF01 (Female)	375	24 dB
	BGM	No.94	6000	

4.3.3　聴取実験による提案法 A の評価

音声と BGM を入力信号組として処理を行い, 聴取実験と考察を行った. 入力信号組は 3 セット用意した. 3 セットは, 音声について内容と話者が, BGM についてジャンルが異なる. 話者は男性 2 名女性 1 名である. 入力信号組の設定を, 表 4.13 [55, 56] に示す. これらの信号は, 前半は音声のみがあり, 後半は音声と BGM の両方があり, それぞれ長さが 6 秒である. 音声と BGM の音量は, 単純加算で音声が聴きとれないように予め調整した. 提案法 A のパラメータは, 3 つの入力信号セットを用いた実験を通して同じ値とした. 実験諸元を表 4.14 に示す.

表 4.14. 提案法 A の実験諸元

Sampling frequency (F_{S})	44100 Hz
Samples of FFT (N_{FFT})	1024
Samples of frame shift	384
Type of analyze window	Hanning
Samples of analyze window	1023
Type of synthesize window	Hanning
Samples of synthesize window	767

図 4.13 に, 提案法 A の特徴が顕著である時間周波数平面を示す. BGM によって埋もれることなく, 音声が聞こえるためには, 音声のスペクトル遷移パターンが, 混合後の時間周波数平面でも見える必要がある. ここで, 単純加算の時間周波数平面 (4.13(c)) では, 音声のスペクトル遷移パターンが見えない. 一方, 提案法 A(4.13(c)) では, 音声のスペクトル遷移パターンが見える. また, BGM のスペクトル遷移パターンも, 提案法 A では明瞭に残っている.

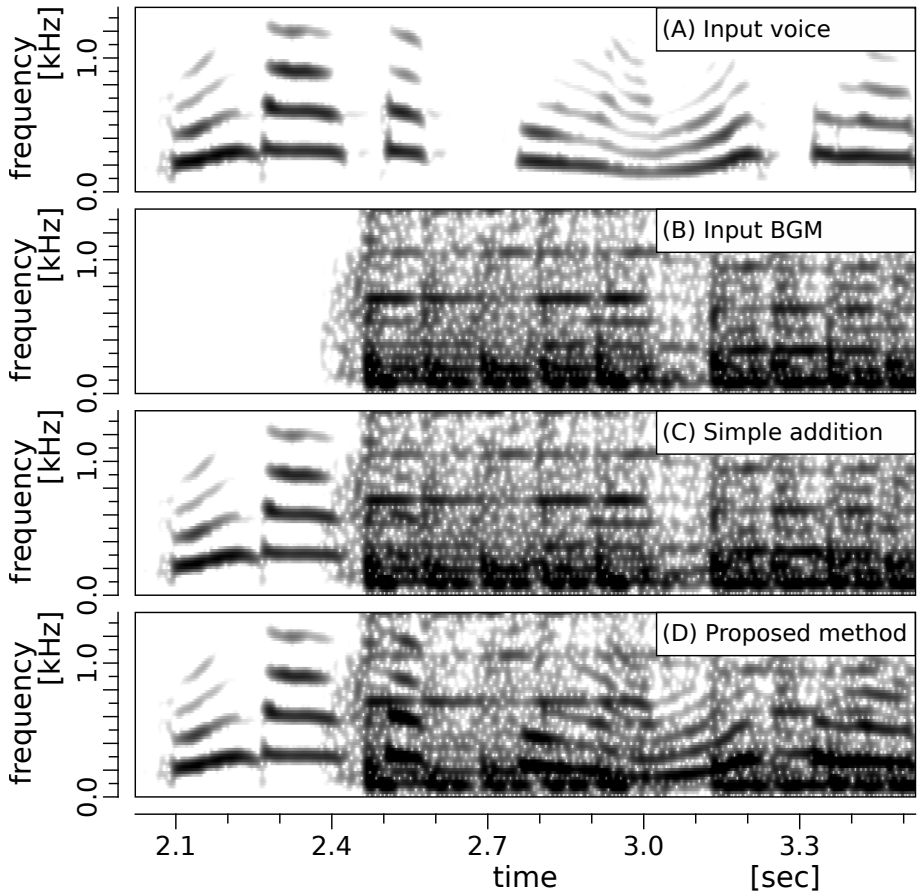


図 4.13. 入力信号をセット 1 としたときの, 入出力信号の時間周波数平面. a,b：入力信号 (a：音声, b: BGM), c.d：出力信号 (c：単純加算, d：提案法 A)

続いて, 3.0∼3.1 sec に着目する. この時区間では, BGM が一時的に無音で, 音声は持続している. この時区間の前後を通じての音声のエネルギーが, 提案法 A では持続している. これはゲイン係数への平滑化の効果の現れである. 単に音量差が大きい時間周波数平面の座標点のみを処理対象とすると, 音声が持続していても BGM に音量変動があるとき, 音声への強調の具合が BGM の音量と連動する. このとき, 音量変動感に聴取の意識が引きずられ, 音声の聴き取りやすさ, 音質が劣化する. 図に示す時区間では, 提案法 A で BGM の音量変動に連動した音量変化による音質劣化を軽減できている.

処理結果について, 聴取実験を行った. 評価対象の音信号は, 9 つのミキシング結果である. 表 4.13 に示す 3 セットを, 単純加算 (B), 従来法として提案法 A の時不変イコライザ近似 (C), 提案法 A(S) で処理した. イコライザ近似の実験諸元を表 4.16 に示す.

提案法 A に対応する従来法の信号 C は, 提案法 A の出力信号と狭帯域スペクトログラムが等しくなるように生成した. 狭帯域スペクトログラムを図 4.14 に示す. 処理後の標準偏差を, 表 4.17 に示す. 続いて, 単純加算と比較しての音量増加量を, 表 4.18 に示す. 提案法 A は, 対応するイコライザ近似よりも単純加算 B に対しての音量変化量が少ない. また, 3 セットでの Perceptual SNR の増加を, 表 4.19 に示す. 提案法 A は, イ

表 4.15. 聴取実験での MOS 値の設定

Point	Impairment
5	Degradation is inaudible
4	Degradation is audible but not annoying
3	Degradation is slightly annoying
2	Degradation is annoying
1	Degradation is very annoying

表 4.16. 提案法 A のイコライザ近似の実験諸元

Sampling frequency (F_{S})	44100 Hz
Samples of FFT (N_{FFT})	128
Samples of frame shift	16
Type of analyze window	Hanning
Samples of analyze window	127
Type of synthesize window	Hanning
Samples of synthesize window	63

コライザ近似よりも Perceptual SNR が小さい.

評価項目は, Q1 : 音声の聞き取りやすさ, Q2 : 音声の音質, Q3 : BGM の音質の 3 つである. これらの項目について, 表 4.15 に示す 5 段階の MOS(mean opinion scores) 値で評価させた. 被験者は 20 代男性 6 人で, 全員にヘッドホンで聴取させた. 聴取実験の結果を, 図 4.15 に示す.

Q1 の結果では, 提案法 A(図 4.15, A) でのみすべてのセットで 3 を超えている. さらに, Q1 での提案法 A とイコライザ近似 (図 4.15, E) の評価には, 有意な差 ($p = 0.027$,

表 4.17. 処理後の標準偏差

入力信号	Set1	Set2	Set3
単純加算	2129.3	4258.6	5932.4
イコライザ近似	2233.9	4611.1	4861.7
提案法 A	2155.3	4213.5	5428.2

表 4.18. 単純加算と比較しての音量増加量

入力信号	Set1 (dB)	Set2 (dB)	Set3 (dB)
イコライザ近似	0.4	0.7	-1.7
提案法 A	0.1	-0.1	-0.8

表 4.19. Perceptual SNR の増加量

入力信号	Set1 (dB)	Set2 (dB)	Set3 (dB)
イコライザ近似	-0.1	0.6	0.7
提案法 A	-0.4	-0.2	0.1

Mann-Whitney test) があった. 対照的に Q2, Q3 では, 提案法 A の評価は低い. 特に, セット 1 での結果は 3 未満である. ここで, このセット 1 の Q1 での評価が十分に高いことに注目する. 従って, 入力信号の他の特性を使用して, 提案法 A の効果の程度を制御することにより, これらの評価の減少を緩和するできるだろう.

さらに, 音声の了解度の聴取実験を行った. 評価音は, 9 つの音声信号,3 の BGM,3 混合方法を組み合わせた. 単語了解度と文章了解度の評価結果を, 表 4.20 に示す. した

表 4.20. 単語了解度, 文章了解度の評価結果

	単純加算 (%)	イコライザ近似 (%)	提案法 A (%)
単語了解度	71.7	83.3	71.7
文章了解度	56.7	76.7	66.7

がって, 音声の聴き取りを維持する手法の有効性が示された.

特筆すべきは set3 である. set3 では, 単純加算で特に音声が聴き取ることができなかったが, 提案法 A では音声の聞き取りができるようになっている. しかも BGM の音質劣化が乏しい. また, 単純加算とイコライザ近似では評価が不快な程度まで下がるが, 提案法 A では不快な程度までは下がらない.

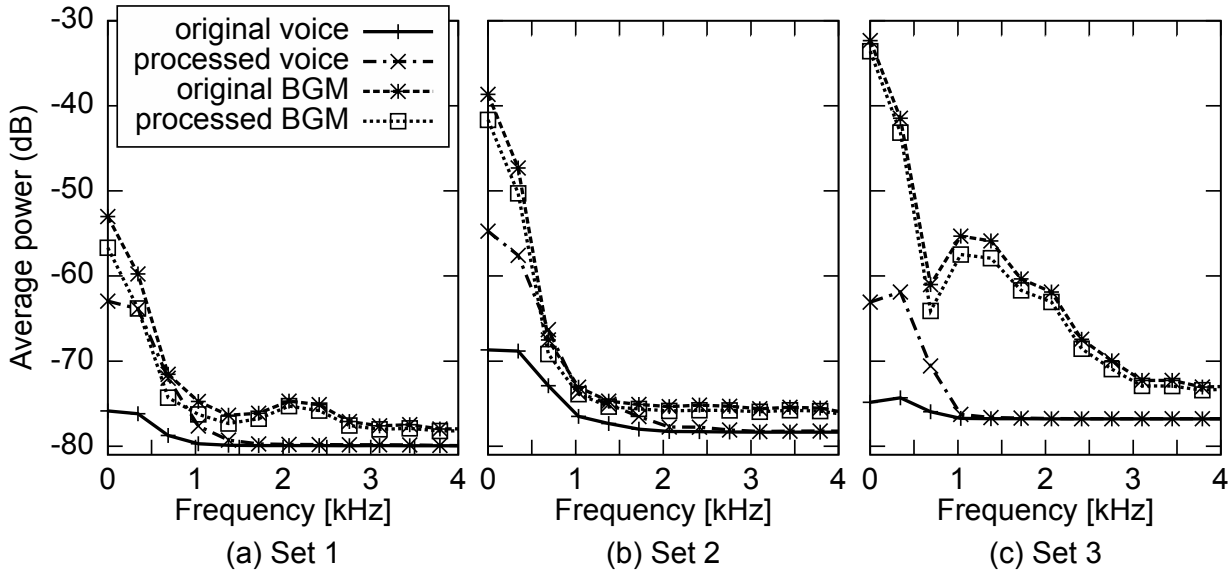


図 4.14. 入力信号, 提案法 A の出力信号の狭帯域スペクトル (帯域幅 = 344.53 Hz, $F_{\mathrm{S}} = 44100$ Hz, $N_{\mathrm{FFT}} = 128$)

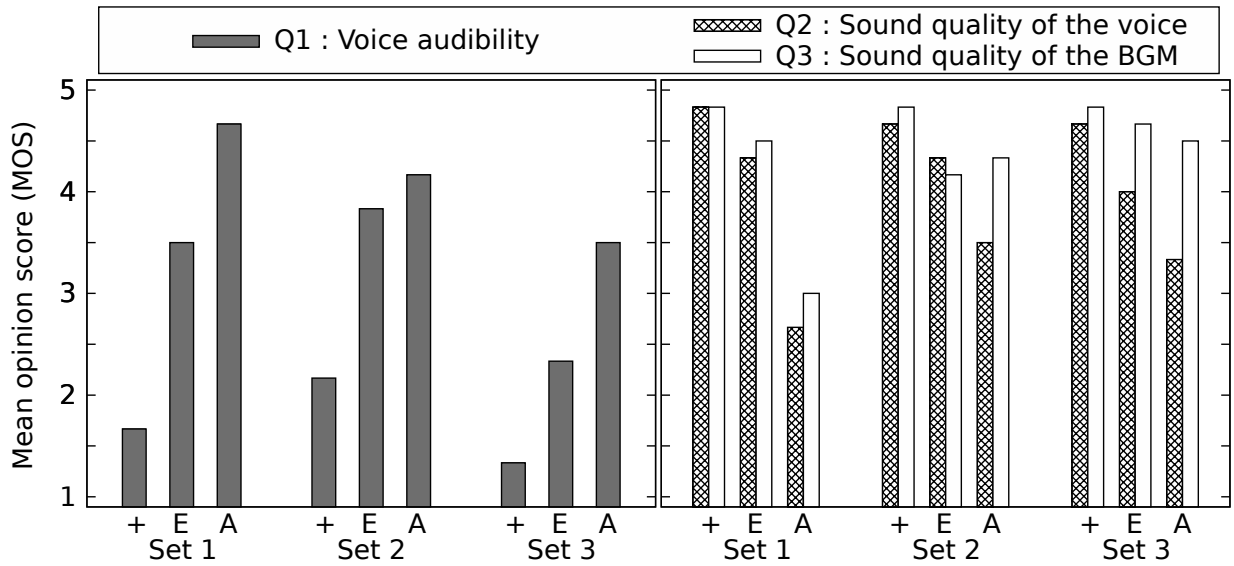


図 4.15. 聴取実験結果. + : 単純加算, E : イコライザ近似, A : 提案法 A

第 5 章
フォルマントに着目した検討

5.1 フォルマント構造に着目した手法

本章では, フォルマント構造 [63, 64, 65] に着目した手法を提案法 B として提案する. フォルマントはスペクトル包絡のピークであり, 母音識別において重要である. 従って, 音声の聴き取りやすさは, 調波構造ではなくフォルマント構造を強調することでも確保できるはずである.

フォルマント構造に着目する利点として, 強調対象の周波数帯域を制限できる点にある. 母音識別に用いられるフォルマントは, 3 kHz より低い周波数帯域に分布 [63, 64] する. 一方, 提案法 A では可聴域に含まれる全ての調波構造を処理対象としている.

また, 低い周波数帯域のみへの処理で聴き取りが確保できれば, サンプリング周波数の低いシステム, 再生帯域の上限が低いスピーカで用いることができる.

5.1.1 フォルマント構造維持を規範とした手法 (提案法 B)

本節では, フォルマント構造維持を規範とした手法である提案法 B について述べる. ブロック図を, 図 5.1 に示す. 処理は以下のように行われる.

1. 入力時間信号 $x_{\mathrm{A}}[n]$(優先音), $x_{\mathrm{B}}[n]$(非優先音) を FFT し, 時間周波数成分 $X_{\mathrm{A}}[i,k]$, $X_{\mathrm{B}}[i,k]$ に展開
2. $X_{\mathrm{A}}[i,k]$, $X_{\mathrm{B}}[i,k]$ を, それぞれパワー $P_{\mathrm{A}}[i,k]$, $P_{\mathrm{B}}[i,k]$ に変換・平滑化
3. $x_{\mathrm{A}}[n]$ に LPC を適用し, スペクトル包絡 $E_{\mathrm{A}}[i,k]$ を抽出

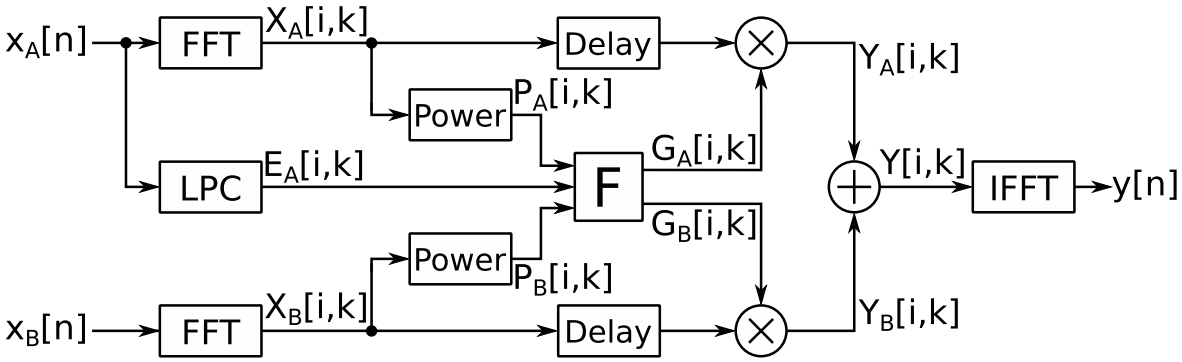


図 5.1. フォルマント構造維持を規範とした音声信号混合法 (提案法 B)

4. $P_{\mathrm{A}}[i,k]$, $P_{\mathrm{B}}[i,k]$, $E_{\mathrm{A}}[i,k]$ より, ゲイン係数 $W_{\mathrm{A}}[i,k]$, $W_{\mathrm{B}}[i,k]$ を決定・平滑化
5. $X_{\mathrm{A}}[i,k]$, $X_{\mathrm{B}}[i,k]$ に $W_{\mathrm{A}}[i,k]$, $W_{\mathrm{B}}[i,k]$ をそれぞれ乗算し, 処理後の入力信号 $Y_{\mathrm{A}}[i,k]$, $Y_{\mathrm{B}}[i,k]$ を生成
6. $Y_{\mathrm{A}}[i,k]$ と $Y_{\mathrm{B}}[i,k]$ を加算し $Y[i,k]$ を生成
7. $Y[i,k]$ を IFFT し, 出力時間信号 $y[n]$ を出力

フォルマント帯域判定 $D_{\mathrm{fn}}[i,k]$ は, 推定されたフォルマント周波数軌跡より決定する. フォルマント帯域幅 w_{fmt} に基づき, 軌跡から周波数方向に上下 $\frac{w_{\mathrm{fmt}}-1}{2}$ ビンをフォルマント帯域とする.

ゲイン操作は, フォルマント帯域のうち音声 BGM ともに有音な領域とする. ゲイン操作実行判定 $D_{\mathrm{G}}[i,k]$ を, 式 (5.1) で定義する. ここで $P_{\mathrm{A}}[i,k]$, $P_{\mathrm{B}}[i,k]$ は優先音および非優先音のパワー, $\mathrm{ATH}[k]$ は最小可聴値 [60] とする.

$$D_{\mathrm{G}}[i,k] = \begin{cases} 1 & \left(\begin{array}{r} \mathrm{ATH}[k] \le P_{\mathrm{A}}[i,k] \\ \wedge \mathrm{ATH}[k] \le P_{\mathrm{B}}[i,k] \\ \wedge D_{\mathrm{fn}}[i,k] \end{array}\right) \\ 0 & (\text{otherwise}) \end{cases} \tag{5.1}$$

$W_{\mathrm{A}}[i,k]$ は, 時間フレーム毎の入力信号のパワー比率 $R_p[i]$, およびパラメータ $p_1 \sim p_4$ に基づき, 式 (5.2) により決定する. ここで, N_{FFT} は FFT 点数である. ゲイン操作実行判定 $D_{\mathrm{G}}[i,k]$ が 1 の場合はゲインを増幅し, 0 の場合はゲインを増幅しない. 式 (5.2) で

作られる $W_{\mathrm{A}}[i,k]$ を, $R_{\mathrm{P}}[i]$ の関数とし, 図 5.2 に表す.

$$W_{\mathrm{A}}[i,k] = \begin{cases} \frac{p_3-1}{p_1-1}R_{\mathrm{P}}[i] + 1 - \frac{p_3-1}{p_1-1} & \left(\begin{array}{r} D_{\mathrm{G}}[i,k]=1 \\ \wedge 1.0 \le R_{\mathrm{P}}[i] < p_1 \end{array}\right) \\ \frac{p_4-p_3}{p_2-p_1}R_{\mathrm{P}}[i] + p_3 - \frac{p_4-p_3}{p_2-p_1} & \left(\begin{array}{r} D_{\mathrm{G}}[i,k]=1 \\ \wedge p_1 \le R_{\mathrm{P}}[i] < p_2 \end{array}\right) \\ p_4 & \left(\begin{array}{r} D_{\mathrm{G}}[i,k]=1 \\ \wedge p_2 \le R_{\mathrm{P}}[i] \end{array}\right) \\ 1.0 & (\text{otherwise}) \end{cases} \tag{5.2}$$

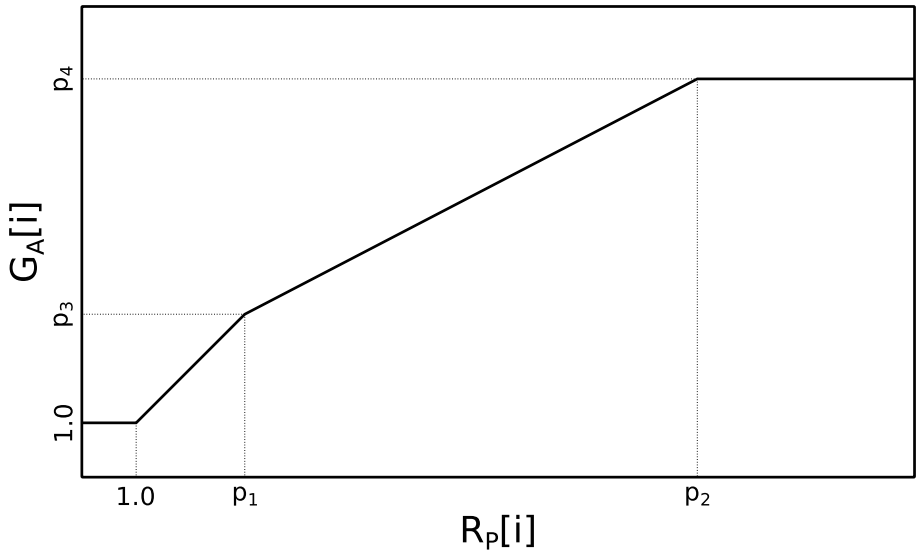


図 5.2. 式 (5.2) における, $R_{\mathrm{P}}[i]$ と $W_{\mathrm{A}}[i,k]$ の関係

$W_{\mathrm{B}}[i,k]$ は, 増幅処理後の優先音のパワーを非優先音のパワーが上回っている場合に, 1 つ前の時間フレーム $i-1$ における同じ周波数帯域のゲイン係数 $W_{\mathrm{B}}[i-1,k]$ を, α 倍ずつ減衰させることで決定する. これは, フレーム間での変化量を固定することで, 急激な音量変化を防ぐ狙いである. もし処理後に $W_{\mathrm{B}}[i,k]$ の下限値 L_α を下回る場合は, α を乗算しないこととする. また, $D_{\mathrm{G}}[i,k]=0$ の場合は $W_{\mathrm{B}}=1$ とする.

$$W_{\mathrm{B}}[i,k] = \begin{cases} \alpha W_{\mathrm{B}}[i-1,k] & \left(\begin{array}{r} D_{\mathrm{G}}[i,k]=1 \\ \wedge P_{\mathrm{A}}[i,k] < P_{\mathrm{B}}[i,k] \\ \wedge L_\alpha < \alpha W_{\mathrm{B}}[i-1,k] \end{array}\right) \\ W_{\mathrm{B}}[i-1,k] & \left(\begin{array}{r} D_{\mathrm{G}}[i,k]=1 \\ \wedge P_{\mathrm{A}}[i,k] < P_{\mathrm{B}}[i,k] \\ \wedge L_\alpha \ge \alpha W_{\mathrm{B}}[i-1,k] \end{array}\right) \\ 1.0 & (\text{otherwise}) \end{cases} \tag{5.3}$$

本稿では, 実験により各パラメータの値を, $p_1=3.0$, $p_2=100.0$, $p_3=3.0$, $p_4=6.0$, $\alpha=0.708$, $L_\alpha=0.01$ とした.

表 5.1. 実験諸元

パラメータ		値
サンプリング周波数	F_{s}	44100 Hz
量子化ビット数	N_{bit}	16 bit
FFT 点数	N_{FFT}	1024 点
解析時窓関数	$H[n]$	ハニング窓形 1024 点
合成時窓関数	$G[n]$	ハニング窓形　256 点
フレームシフト	N_{SFT}	256 点
LPC 次数	N_{LPC}	12 次
フォルマント帯域幅	W_{fmt}	11 ビン (473.73 Hz)

5.1.2　聴取実験による提案法 B の評価

評価対象について述べる. 入力信号は, 表 5.2 に示す S1∼S3 の 3 種類の音源セットとした. 音声には SRV-DB[55] の「発話のプロフェッショナルによる編集手帳（読売新聞）の読み上げ」のデータを, 音楽には「RWC 研究用音楽データベース：ポピュラー音楽 [56]」のデータを用いた. 今回の実験では, 非優先音の音量を優先音の音量より, 表 5.2 に示す音量差分あらかじめ大きく設定している. 各セットをこの音量差で単純加算したとき, 音声は音楽に埋もれ, 聴き取りが困難である. 被験者は成人男性 8 名であり, 全員ヘッドフォンによる聴取を行った.

次に, ミキシング手法は表 5.3 に示す 5 種類とした. 参考のために, 調波構造に着目した手法 (手法 8(第 4.2 節), C), 手法 8 について処理領域をフォルマント帯域に制限した手法 ($\mathrm{C_f}$), 提案法 B について処理領域をフォルマント帯域に制限せず音声 BGM ともに有音な領域とした手法 (P) とも比較した.

評価項目は, 音声の聴き取りやすさ, 音声の自然さ, 音楽の自然さの 3 項目とし, 5 段階の MOS 値 (1：Bad, 2：Poor, 3：Fair, 4：Good, 5：Excellent) による評価を行った.

実験諸元を表 5.1 に示す.

表 5.2. 音源セット

セット名	音声 (優先音)	音楽 (非優先音)	音量差
S1	女性音声	J-POP	12 dB
S2	男性音声	シティーロック	12 dB
S3	女性音声	ポップス	24 dB

表 5.3. 評価対象

音源	備考
B	単純加算
C	調波構造に着目した手法 (手法 8)
C_f	手法 8 の処理範囲をフォルマント帯域に制限した手法
P	提案法 B の処理範囲をフォルマント帯域に制限しない手法
P_f	フォルマント構造維持を規範とした手法 (提案法 B)

聴取実験結果について考察する. 図 5.3 に各評価項目ごとの平均 MOS 値を示す.

P と P_f を比較する. S1 と S2 においては, 処理領域の制限により音声と音楽の自然さが改善され, かつ音声の聴き取りやすさが維持されている. S3 でも, 音楽の音質の自然さが改善している. この結果より, フォルマント帯域への処理領域の制限が有効であると示された.

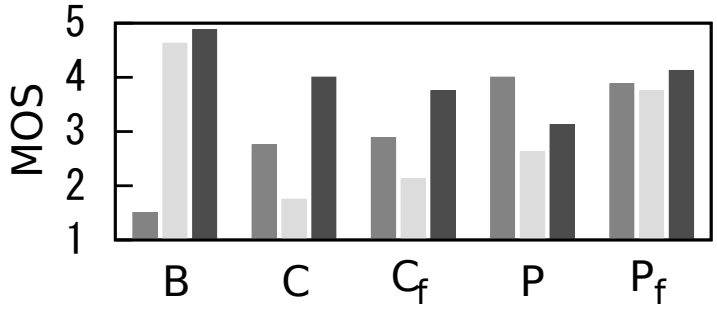
C と C_f を比較する. S2 の音声の聴きとりやすさ, S1 と S3 の音声の音質の自然さが向上している. 一方, S3 の音声の聴きとりやすさ, S1 と S3 の BGM の音質の自然さが劣化している. 手法 8 へのフォルマント帯域への処理領域の制限では, 効果が限定的である.

C と P_f を比較する. S1 と S2 の BGM の音質の自然さが同等, S2 の音声の音質の自然さで C が良好である以外は, P_f が良好である.

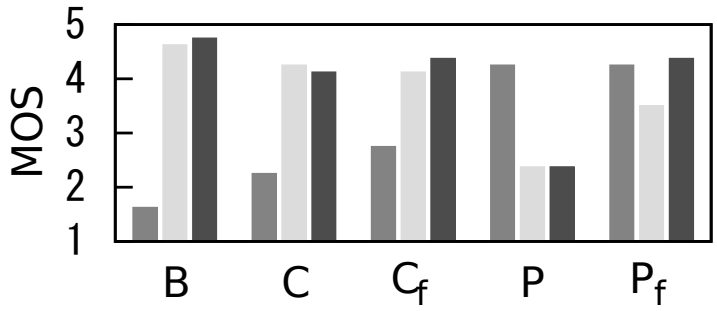
表 5.4. 日本語子音の分類 [63]

調音位置		口唇		歯, 歯茎		口蓋		声門
音源		有声	無声	有声	無声	有声	無声	無声
調音方式	摩擦音		f	z	s	ʒ	ʃ	h
	破擦音			dz	ts	dʒ	tʃ	
	破裂音	b	p	d	t	g	k	
	半母音	w		r		j		
	鼻音	m		n		ŋ		

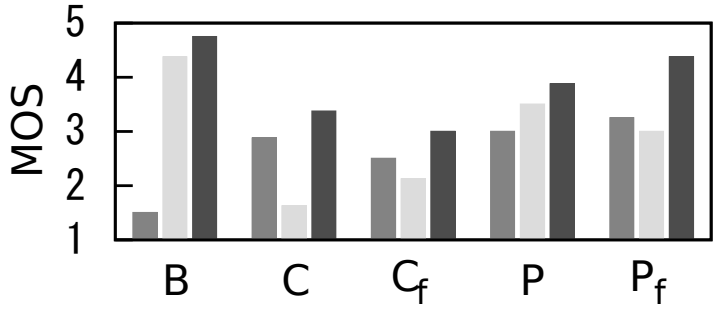
各音源について, 聴き取りづらい時区間について時間周波数平面の観察を行った. 観察により, 子音『t』, 『k』, 『z』, 『s』, 『ts』の音節に該当することがわかった. 表 5.4 に日本語子音の分類を示す. 問題に該当する音節の子音は, 摩擦音や破擦音, 破裂音



(a) S1



(b) S2



(c) S3

図 5.3. 聴取実験による評価項目ごとの平均 MOS 値. B, C, C_f, P, P_f は, ミキシング手法であり, 表 5.3 に示した. 評価項目はミキシング手法ごとに左から, 音声の聴き取りやすさ, 音声の自然さ, 音楽の自然さ

の子音に含まれていることが分かる. 特に歯茎無声音は, すべての子音が該当している. 従って, これらの子音が現れた場合, 提案法 B の処理では埋もれを回避することができないと考えられる.

第 6 章
フォルマント構造と歯擦音に着目した検討

本章では, 音声と BGM の同時再生時に音声が埋もれない音声信号混合法の構成法を, 提案法 C として提案する.

音声が埋もれない音声信号混合法として, 調波構造に着目した手法 (提案法 A) とフォルマント構造に着目した手法 (提案法 B) を提案してきた. このうち, 提案法 B について, 特に歯擦音の聴きとりやすさが不得意であると主観評価によってわかっている (5.1.2 節). そこで本章では, 提案法 B をベースとして歯擦音への識別性を向上させる手法を提案する.

提案法 B では, 母音認識に重要なフォルマントが主として分布する, 低い周波数帯域のフォルマント帯域のみに処理を限定していた. 一方, 歯擦音の成分は, それらの周波数帯域より高い周波数帯域に分布しており, 提案法 B では対応できない.

そこで提案法 C では, 提案法 B で処理対象外とする周波数帯域より高い周波数帯域に対して, 歯擦音のスペクトルを強調する処理を, 提案法 B に追加する.

なお, 提案法 A では歯擦音を含む非純音のスペクトルを強調処理も行なっている. このため, 提案法 A について歯擦音のスペクトルを強調する処理を新たに追加する検討は不必要である.

6.1 節で, 提案法 C として歯擦音への処理を提案法 B に追加した手法を提案する. 6.2 節で, 実信号を用いて検討する.

表 6.1. 実験諸元

パラメータ		値
サンプリング周波数	F_{s}	44100 Hz
量子化ビット数	N_{bit}	16 bit
FFT 点数	N_{FFT}	1024 点
解析時窓関数	$H[n]$	ハニング窓形 1024 点
合成時窓関数	$G[n]$	ハニング窓形　256 点
フレームシフト	N_{SFT}	256 点
LPC 次数	N_{LPC}	12 次
フォルマント帯域幅	W_{fmt}	11 ビン (473.73 Hz)

6.1　歯擦音と母音の識別性を重視する音声信号混合法 (提案法 C)

まず, 提案法 C の元となるフォルマント構造に着目した手法 (提案法 B) について, 提案法 C で追加する歯擦音への処理との対比が容易となるように立式し直す.

入力音声信号, 入力 BGM 信号を, 短時間フーリエ変換により, それぞれ複素時間周波数表現 $X_{\mathrm{A}}[i,k]$, $X_{\mathrm{B}}[i,k]$ に展開する. ここで, i をフレーム番号, k を周波数ビン番号とする. フォルマント周波数推定, 処理対象のフォルマント帯域の決定を経て, 両信号へのゲイン係数 $W_{\mathrm{A}}[i,k]$, $W_{\mathrm{B}}[i,k]$ を算出する. $X_{\mathrm{A}}[i,k]$, $X_{\mathrm{B}}[i,k]$ へ $W_{\mathrm{A}}[i,k]$, $W_{\mathrm{B}}[i,k]$ をそれぞれ適用し, それらどうしの加算により, 出力信号の複素時間周波数表現 $Y[i,k]$ を得る. $Y[i,k]$ の逆短時間フーリエ変換により, 出力信号を算出する. 時間周波数表現の実験諸元を, 表 6.1 に示す.

まず, 音声のフォルマント周波数を推定する. 本手法では, 線形予測符号 (Linear Predictive Coding, 以下 LPC)[63, 64] による推定法を用いる.

LPC は, 式 (6.1) で定義される.

$$y'_n = -\sum_{p=1}^{N_{\mathrm{LPC}}} a_p y_{n-p} \tag{6.1}$$

y_n, y'_n は, それぞれ n サンプルでの実測値と予測値であり, a_p は LPC 係数, N_{LPC} は LPC 次数である.

LPC 係数 a_p と残差分散 $E(=\sigma^2)$ から, 周波数特性 $f(\omega)$ が式 (6.2) として得られる.

得られた $f(\omega)$ のピークを探すことで, フォルマント周波数が推定できる.

$$f(\omega) = \frac{\sigma^2}{2\pi} \frac{1}{|1 + \sum_{p=1}^{N_{\mathrm{LPC}}} a_p e^{-jp\omega}|^2} \tag{6.2}$$

処理対象のフォルマント帯域 $D_{\mathrm{fn}}[i,k]$ (Boolean) は, LPC によって推定されたフォルマント周波数軌跡より決定する. 本手法では, フォルマント軌跡から周波数方向に, 上下 $\frac{w_{\mathrm{fmt}}-1}{2}$ ビンに拡げる. また, 処理対象のフォルマント帯域は, $50 \sim 6,000$ Hz の周波数ビンに制限する.

フォルマント帯域 $D_{\mathrm{fn}}[i,k]$ に従い, ゲイン係数 $W_{\mathrm{A}}[i,k]$, $W_{\mathrm{B}}[i,k]$ を生成する.

$D_{\mathrm{G}}[i,k]$ を式 (6.3) で定義する. ここで $P_{\mathrm{A}}[i,k]$, $P_{\mathrm{B}}[i,k]$ は, 複素数 $X_{\mathrm{A}}[i,k]$, $X_{\mathrm{B}}[i,k]$ のパワー, $\mathrm{ATH}[k]$ は最小可聴値 [59] である.

$$D_{\mathrm{G}}[i,k] = \begin{cases} 1 & \left(\begin{array}{r} \mathrm{ATH}[k] \leq P_{\mathrm{A}}[i,k] \\ \wedge \mathrm{ATH}[k] \leq P_{\mathrm{B}}[i,k] \\ \wedge D_{\mathrm{fn}}[i,k] \end{array} \right) \\ 0 & (\text{otherwise}) \end{cases} \tag{6.3}$$

音声信号へのゲイン係数 $W_{\mathrm{A}}[i,k]$ は, 式 (6.4~6.6) に示すように算出する. 時間フレーム i ごとの両入力信号間のパワー比率 $R_{\mathrm{p}}[i]$ を元に, フレームごとのゲイン $W_{\mathrm{p}}[i]$ を, パラメータ $p_1 \sim p_4$ で形作られる非線形関数を用いて算出する. $W_{\mathrm{p}}[i]$ は, 処理対象のフォルマント帯域 $D_{\mathrm{G}}[i,k]$ に該当する周波数ビンにのみ適用する.

$$R_{\mathrm{p}}[i] = \frac{\sum_{k=0}^{N_{\mathrm{FFT}}} P_{\mathrm{B}}[i,k]}{\sum_{k=0}^{N_{\mathrm{FFT}}} P_{\mathrm{A}}[i,k]} \tag{6.4}$$

$$W_{\mathrm{p}}[i] = \begin{cases} \frac{p_3-1}{p_1-1} R_{\mathrm{P}}[i] + 1 - \frac{p_3-1}{p_1-1} & \left(1.0 \leq R_{\mathrm{p}}[i] < p_1 \right) \\ \frac{p_4-p_3}{p_2-p_1} R_{\mathrm{P}}[i] + p_3 - \frac{p_4-p_3}{p_2-p_1} & \left(p_1 \leq R_{\mathrm{p}}[i] < p_2 \right) \\ p_4 & \left(p_2 \leq R_{\mathrm{p}}[i] \right) \\ 1.0 & (\text{otherwise}) \end{cases} \tag{6.5}$$

$$W_{\mathrm{A}}[i,k] = \begin{cases} W_{\mathrm{p}}[i] & \left(D_{\mathrm{G}}[i,k] = 1 \right) \\ 1.0 & (\text{otherwise}) \end{cases} \tag{6.6}$$

BGM 信号へのゲイン係数 $W_{\mathrm{B}}[i,k]$ は, その周波数ビン k での, 音声のパワー $P_{\mathrm{A}}[i,k]$ より BGM のパワー $P_{\mathrm{B}}[i,k]$ が大きいとき, 1 つ前のフレームにおけるゲイン係数 $W_{\mathrm{B}}[i-1,k]$

を, Δ_{B} 倍ずつ減衰させる. この機構により, BGM への急激な音量減衰が抑えられる. 加えて, 減衰の累積による大幅な音量減衰への対処として, 下限値 L_α を設けている.

$$W_{\mathrm{B}}[i,k] = \begin{cases} \Delta_{\mathrm{B}} W_{\mathrm{B}}[i-1,k] & \left(\begin{array}{r} D_{\mathrm{G}}[i,k] = 1 \\ \wedge R_{\mathrm{p}}[i] > 1 \\ \wedge L_\alpha \leq \Delta_{\mathrm{B}} W_{\mathrm{B}}[i-1,k] \end{array} \right) \\ W_{\mathrm{B}}[i-1,k] & \left(\begin{array}{r} D_{\mathrm{G}}[i,k] = 1 \\ \wedge R_{\mathrm{p}}[i] > 1 \\ \wedge L_\alpha > \Delta_{\mathrm{B}} W_{\mathrm{B}}[i-1,k] \end{array} \right) \\ 1.0 & (\text{otherwise}) \end{cases} \tag{6.7}$$

本稿では各パラメータの値を, $p_1 = 3.0$, $p_2 = 200.0$, $p_3 = 3.0$, $p_4 = 6.0$, $\Delta_{\mathrm{B}} = 0.708$, $L_\alpha = 0.1$ とした. さらに, ゲイン係数 $W_{\mathrm{A}}[i,k]$, $W_{\mathrm{B}}[i,k]$ は, ゲイン適用前に LPF によって時間方向に平滑化する. 本稿では LPF を, 1 次 IIR フィルタによる指数平滑化法とし, 時定数を $\tau_{\mathrm{s}} = 20$ msec とする.

6.1.1　歯擦音に特化したゲイン係数

提案法 C では, 新たに歯擦音に特化したゲイン係数を生成し, 提案法 B のゲイン係数と重ね合わせる. これにより, 歯擦音と母音の識別性を両立を図る.

歯擦音のエネルギは, 成人男性では 4 kHz 以上の周波数帯域に存在し, 女性や子供ではより高い周波数帯域に存在 [65] する. 一方, 母音認識で重要なフォルマントは, 主として 3 kHz より低い周波数帯域に分布 [63, 64] する. 母音と歯擦音とでは, 重要な周波数帯域は異なるものと見なせる. 母音認識特化と歯擦音特化のゲイン係数の重ね合わせは, 特定の周波数帯域を境とした単純な方法でも, 効果が期待できる.

歯擦音の時間周波数平面上での領域として, フォルマント帯域よりも高い周波数帯域で, 有音かつ純音でない領域とする. 純音判定 $D_{\mathrm{tm}}[i,k]$ は, MPEG-1 audio の Psychoacoustic Model 1[59] に含まれる判定を用いる.

提案法 C では, 歯擦音に特化したゲイン係数を, 提案法 B のゲイン係数生成法である式 (6.4∼6.7) を基に, 式 (6.8∼6.12) にように生成する. ここで, 提案法 B から変更点の

ある変数を $\cdot^{\mathrm{S}}$ で区別する.

$$D^{\mathrm{S}}_{\mathrm{G}}[i,k] = \begin{cases} 1 & \left(\begin{array}{r} s_1\mathrm{ATH}[k] \le P_{\mathrm{A}}[i,k] \\ \wedge s_1\mathrm{ATH}[k] \le P_{\mathrm{B}}[i,k] \\ \wedge \neg D_{\mathrm{tm}}[i,k] \\ \wedge \neg D_{\mathrm{tm}}[i,k-1] \\ \wedge \neg D_{\mathrm{tm}}[i,k+1] \\ \wedge P_{\mathrm{A}}[i,k] < P_{\mathrm{B}}[i,k] \end{array} \right) \\ 0 & (\text{otherwise}) \end{cases} \tag{6.8}$$

$$W^{\mathrm{S}}_{\mathrm{A}}[i,k] = \begin{cases} s_2 W_{\mathrm{p}}[i] & \left(D^{\mathrm{S}}_{\mathrm{G}}[i,k] = 1 \right) \\ 1.0 & (\text{otherwise}) \end{cases} \tag{6.9}$$

$$W^{\mathrm{S}}_{\mathrm{B}}[i,k] = \begin{cases} \Delta_{\mathrm{B}} W^{\mathrm{S}}_{\mathrm{B}}[i-1,k] & \left(\begin{array}{r} D^{\mathrm{S}}_{\mathrm{G}}[i,k] = 1 \\ \wedge R_{\mathrm{p}}[i] \cdot L_\alpha > 1 \\ \wedge L_\alpha \le \Delta_{\mathrm{B}} W_{\mathrm{B}}[i-1,k] \end{array} \right) \\ W^{\mathrm{S}}_{\mathrm{B}}[i-1,k] & \left(\begin{array}{r} D^{\mathrm{S}}_{\mathrm{G}}[i,k] = 1 \\ \wedge R_{\mathrm{p}}[i] \cdot L_\alpha > 1 \\ \wedge L_\alpha > \Delta_{\mathrm{B}} W_{\mathrm{B}}[i-1,k] \end{array} \right) \\ 1.0 & (\text{otherwise}) \end{cases} \tag{6.10}$$

歯擦音への処理を提案法 B での処理 (式 (6.4~6.7)) として, 実信号で検討したところ, BGM の抑制は過度となり, 一方では音声の強調が十分に感じられず, 効果的でなかった.

そこで, 提案法 C では, BGM の過度な抑制に対して, 有音判定の閾値をパラメータ s_1 によって調整できるようにし, 処理対象とする成分を絞る. さらに, 許容限 L_α を上回る抑制が起こりうる音量差となる領域を, 処理対象から外す. また, 音声への強調をパラメータ s_2 によって調整する. この 3 点の変更により, BGM の音質劣化は緩和され, 歯擦音の聴きとりやすさが得られた. 本稿では各パラメータの値を, $s_1 = 4.0$, $s_2 = 2.0$ とした.

提案法 B によって生成される母音特化のゲイン係数を, $W^{\mathrm{F}}_{\mathrm{A}}$, $W^{\mathrm{F}}_{\mathrm{B}}$ とし, 提案法 C での W_{A}, W_{B} を, $W^{\mathrm{F}}_{\mathrm{A}}$ と $W^{\mathrm{S}}_{\mathrm{A}}$, $W^{\mathrm{F}}_{\mathrm{B}}$ と $W^{\mathrm{S}}_{\mathrm{B}}$ とを, 式 (6.11, 6.12) に示すように, 周波数ビン k_{S} を

表 6.2. 実験で用いた音声信号と, その内容

Symbols	Numbers in the Database[55]	Sentence
Vk	PF02 67	これが危機である.
Vs	PF02 68	これが四季である.
Vz	PF02 71	これが時期である.

境として組み合わせて生成する.

$$W_{\mathrm{A}}[i,k] = \begin{cases} W_{\mathrm{A}}^{\mathrm{F}}[i,k] & \left(k \leq k_{\mathrm{S}} \right) \\ W_{\mathrm{A}}^{\mathrm{S}}[i,k] & \left(k > k_{\mathrm{S}} \right) \end{cases} \tag{6.11}$$

$$W_{\mathrm{B}}[i,k] = \begin{cases} W_{\mathrm{B}}^{\mathrm{F}}[i,k] & \left(k \leq k_{\mathrm{S}} \right) \\ W_{\mathrm{B}}^{\mathrm{S}}[i,k] & \left(k > k_{\mathrm{S}} \right) \end{cases} \tag{6.12}$$

本稿では, 実信号での検討から k_{S} を 4 kHz 相当とした.

6.2　聴取実験による提案法 C の評価

提案法 C を実信号を用いて検討する. 音声には, 「話速バリエーション型音声データベース SRV-DB」[55] に含まれる, 「2. 発話のプロフェッショナルによるオリジナル原稿 (カーナビ文章) の読み上げ」から, 表 6.2 に示す 3 音源を用いた. 発話者は, 女性アナウンサーである. BGM には, 「 RWC 研究用音楽データベース:ポピュラー音楽 [56] 」の楽曲を用いた. 音声と BGM の音量差は, 単純加算時に音声が埋もれるよう調整した. 音声, BGM それぞれの有音な時区間での標準偏差は, 500, 4000 とし, 音声に対し BGM が 18 dB 大きい設定となった. 入力信号と提案法 B および提案法 C の処理結果の時間周波数平面を, 図 6.1 に示す.

入力音声信号 Vk, Vs, Vz の時間周波数平面が, それぞれ図 6.1([k,s,z],v) である. 表 6.2 より, 3 信号は 1 つの子音のみが異なる組で, それぞれ音素は Vk が/k/, Vs が/s/, Vz が/z/, 共通する母音の音素は/i/である.

図 6.1. 入力信号と，単純加算，提案法 B，提案法 C での出力信号の時間周波数平面．行方向に同一の音声信号，列方向に同一のミキシング手法での出力信号を配置した．(k,·) は音声 Vk，(i,·) は音声 Vi，(z,·) は音声 Vz の入出力信号である．(x,·) は入力 BGM 信号そのもの（音声が無音相当）であり，3 画像は同一である．(·,v) は入力音声信号，(·,A) は単純加算の出力信号，(·,F) は提案法 B の出力信号，(·,S) は提案法 C の出力信号である．(k,·)，(i,·)，(z,·) の 3 列の時間周波数平面には下部に，その時区間での発話内容を付した

子音が異なる時区間は, 1.6∼1.8 秒である. /s/では 2∼14 kHz, /z/では 5∼10 kHz の 1 帯域に, エネルギが集中して分布している. /k/では 0∼14 kHz のうち, 数帯域に分かれて分布している. また, 開始部でのエネルギ増加が, /s/, /z/は時間方向に緩やかである. 対して/k/では急峻であり, 開始直後に 0.01 秒程度の無音区間がみられる. これらの特徴は, 単純加算 (図 6.1([k,s,z],A)) と提案法 B(図 6.1([k,s,z],F)) では, BGM の分布 (図 6.1(x,·)) に埋もれている. 提案法 C(図 6.1([k,s,z],S)) では, 該当する領域のエネルギが強調されている.

また, 図 6.1([k,s,z],v) では, エネルギの大きいフォルマント周波数が 3 成分みられ, 0∼4 kHz に存在する. これらは, 低い周波数帯域のものから, F1, F2, F3 である. F1, F2 は, これらが分布する周波数帯域の組み合わせから, 母音認識が行われている. F3 以上の帯域は, 話者識別に影響 [66] があることが示されている. フォルマント周波数の分布の模様は, 単純加算 (図 6.1([k,s,z],A)) では埋もれており, 提案法 B(図 6.1([k,s,z],F)) では, フォルマント周波数間のエネルギが疎な領域の分布で顕著である. 提案法 C(図 6.1([k,s,z],S)) では, 提案法 B(図 6.1([k,s,z],F)) と同様のエネルギ分布がみられる.

提案法 C による歯擦音への考慮が, 母音への強調を阻害せず, 両処理の重ね合わせが効果的に実施されていることが示された.

最後に著者が主観評価によって, 提案法 B と提案法 C を比較した. 比較の結果, 提案法 B では聴き分けができなかった子音について, 提案法 C では明瞭に聴き分けることができることを確認した.

第 7 章
提案法 A, B, C の比較検討

本論文ではここまでに, 音声を埋もれさせない音信号混合法のスマートミキサーでの実装法について, 三手法提案した. 提案法 A は, 調波構造の維持に着目した手法であり, 4 章で述べた. 提案法 B は, フォルマント構造に着目した手法で, 5 章で述べた. 提案法 C は, フォルマント構造と歯擦音の周波数分布に着目した手法で, 6 章で述べた.

本章では, 提案した三手法を比較検討する. 表 7.1 に提案した三手法での, 聴覚心理モデルおよびフォルマント推定の利用の有無をまとめた. ここで, 提案法 C でも聴覚心理モデルを用いている点に留意する.

表 7.1. 本論文で提案した三手法での聴覚心理モデルおよびフォルマント推定の利用の有無

	提案法 A	提案法 B	提案法 C
聴覚心理モデル	○	×	○
フォルマント推定	×	○	○

7.1 実行時間による比較検討

提案した 3 手法について, 6 秒の音データを 10 回処理し, 実行時間の平均値を算出した. 用いた計算機は, CPU が Intel Xeon W3520 2.67GHz で, メモリが 12GB である.

算出結果を表 7.2 に示す. 表 7.2 より, 提案法 A が最も計算量が軽く, 再生時間の 1/6 倍の時間で処理できることがわかった. 従って, 提案法 A はリアルタイムでの処理が可能である. 一方, 提案法 B と提案法 C は同等に計算量が大きく, 処理に再生時間の 2 倍以上の時間が必要である.

表 7.2. 提案した 3 手法の平均実行時間

手法	平均実行時間 (s)
提案法 A	0.50
提案法 B	14.23
提案法 C	14.24

表 7.3. 主観評価での, DMOS 値の設定

評点	基準
5	劣化が全く認められない
4	劣化が認められるが気にならない
3	劣化がわずかに気になる
2	劣化が気になる
1	劣化が非常に気になる

7.2　聴取実験による比較検討

主観評価と発話内容の了解度の聴取実験を実施した.

主観評価は, Q1 : “入力音声と比較して, 内容の聴きとりやすさ”, Q2 : “入力音声と比較して, 音質の自然さ”, Q3 : “入力 BGM と比較して, 音質の自然さ” の 3 つの評価項目とした. 評価値は, 表 7.3 に示す 5 段階の DMOS(Degradation Mean Opinion Score) 値 [67] とした. ここで, 3 つの主観評価項目の研究目的での重要度として, Q1 で高得点となることが最重要であり, Q2 と Q3 では高得点が両立されることが重要であるとする.

発話内容の了解度の評価は, 一度の聴取で聴き取れた内容をひらがなで記述させた. 了解度の算出は, ひらがな一文字単位で正誤を判定し, 各文について, 音節単位, 一文字単位, 一文字違い部のみの三方法で集計した.

音声には,「話速バリエーション型音声データベース SRV-DB」[55] に含まれる,「1. 発話のプロフェッショナルによるオリジナル原稿 (一文字違い文章) の読み上げ」から, 表 7.4 に示す 20 音源を用いた. 発話者は, 女性アナウンサーである. BGM には,「 RWC 研究用音楽データベース:ポピュラー音楽 [56] 」の楽曲を用いた. 音声と BGM の音量差は, 単純加算時に音声が埋もれるよう調整した. 音声, BGM それぞれの有音な時区間での標準偏差は, 500, 4000 とし, 音声に対し BGM が 18 dB 大きい設定となった. 発話内容は, 一文字で共通の母音で子音のみが異なる 4 組から成る.

表 7.4. 実験で用いた音声信号と, その内容

Numbers in the Database[55]	Sentence
PF02 33	彼は貝にこだわる.
PF02 38	彼は差異にこだわる.
PF02 39	彼は鯛にこだわる.
PF02 34	彼は紀伊にこだわる.
PF02 40	彼は地位にこだわる.
PF02 41	彼は二位にこだわる.
PF02 54	その仮面は傾いている.
PF02 59	その画面は傾いている.
PF02 60	その座面は傾いている.
PF02 62	その場面は傾いている.
PF02 53	そのお面は傾いている.
PF02 55	その湖面は傾いている.
PF02 57	その木綿は傾いている.
PF02 58	その路面は傾いている.
PF02 67	これが危機である.
PF02 68	これが四季である.
PF02 71	これが時期である.
PF02 65	これが駅である.
PF02 69	これが席である.
PF02 70	これが劇である.

表 7.5. 聴取実験で比較する手法

記号	手法
+	単純加算
A	提案法 A(調波構造に立脚した提案法
B	提案法 B(フォルマント帯域のみを強調する提案法)
C	提案法 C(フォルマントに加えて歯擦音も強調する提案法)

処理内容は, 表 7.5 に示す 4 手法である.

被験者は 20 代男性 8 名で, 全員ヘッドホンで聴取した. 各被験者に提示する音源のセットについては, 同一文章の提示が 1 回となるように分配した. 比較する手法が 4 手法であることから, 1 つの評価対象の音源について 2 人分の評価結果を得た.

評価結果を, 図 7.1 に示す. (c) 一文字違い部での正解率では, 子音の分類別 (All : 全て, [s z] : 歯擦音, [k g t b] : 撥音, [a n m r] : 母音および無声音) での結果を併記した. 子音の分類別での文章数は, All : 20 文, [s z] : 5 文, [k g t b] : 10 文, [a n m r] : 5 文である. また, 聴取実験結果へのウィルコクソンの順位和検定の両側検定における p 値を, 表 7.6 に示す.

単純加算に対しては, Q3 以外の全ての項目について, 3 手法ともで上回った. 発話内容の了解度は, 単純加算では 5 割程度であったのに対し, 3 手法ともで 9 割程度を達成している. 特に, 提案法 C が最良で, 95%を超えている.

主観評価では, Q3 での提案法 A を除いて, 4 と 3 の間の評価を得ており, 劣化が許容できる程度に留められている.

これらより, 提案法の第一目的である内容把握については, 3 手法ともで達成できることが示された. 加えて, 内容把握と音質の両立については, 提案法 B と提案法 C で達成できることが示された. 3 手法の中では提案法 C が最良であった.

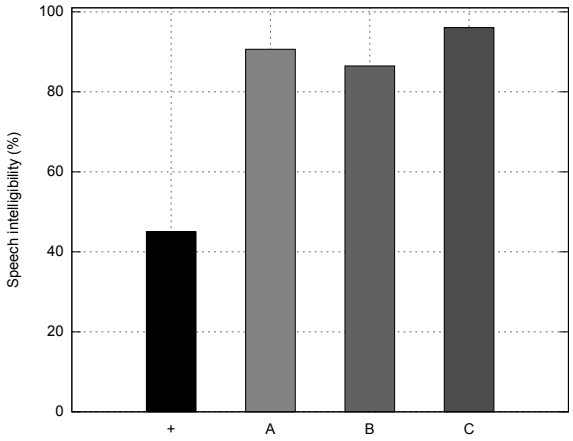
評価結果について考察する.

まず, 提案法 B と提案法 C を比較する. 提案法 C は, 提案法 B に歯擦音相応の成分への処理を追加した手法である.

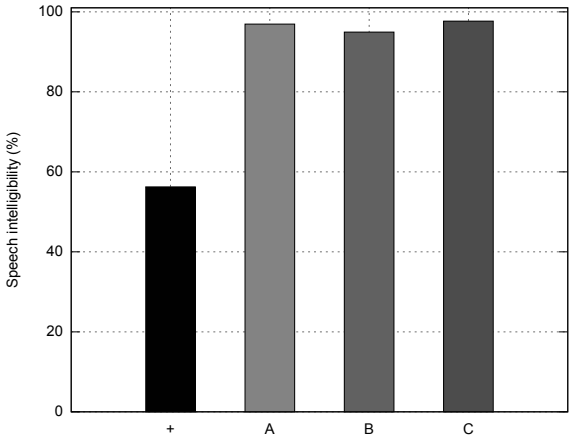
提案法 C で主眼としている歯擦音の聴き取りについて, (c) 一文字違い部での正解率の “[s z] : 歯擦音” で, 30%から 100%へと改善できている. 一方, 歯擦音以外の “[k g t b] : 撥音”, “[a n m r] : 母音および無声音” での正解率は, 5∼10%の改悪となっている. 原因として, 歯擦音相応の成分の強調によって, 他の成分への認知が相対的に低下したと考えられる.

主観評価では, 評価が改善されているか拮抗している. 提案法 C での処理の追加が効果的であることが示された.

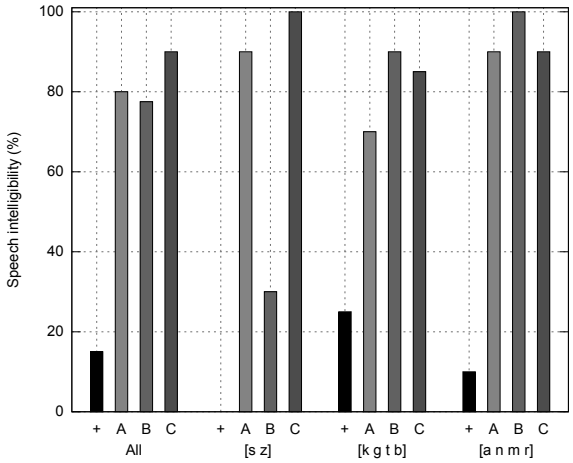
続いて, 調波構造の維持に着目した手法 (提案法 A) とフォルマント構造に着目した手法 (提案法 B および提案法 C) とを比較する. 提案法 A と提案法 B および提案法 C とでは, 時間周波数平面上で処理を施す成分が前者で多い. 前者では有音な成分の全てを



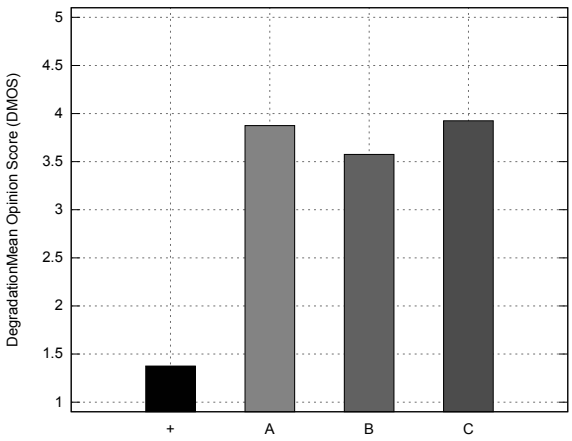
(a) 音節単位での正解率



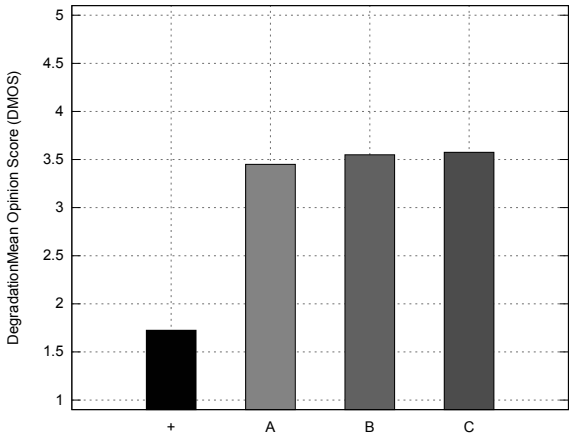
(b) 一文字単位での正解率



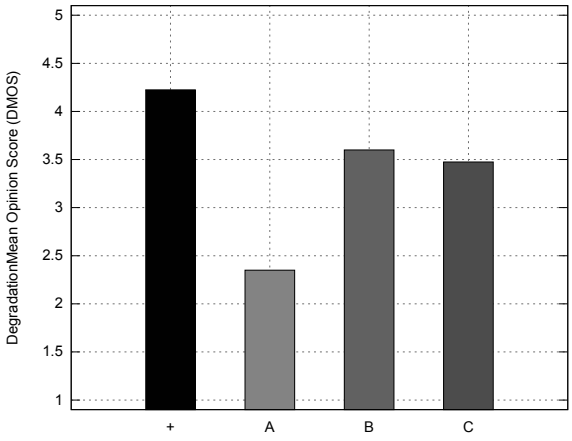
(c) 一文字違い部での正解率



(d) Q1： 入力音声と比較して, 内容の聴きとりやすさ



(e) Q2：入力音声と比較して, 音質の自然さ



(f) Q3：入力 BGM と比較して, 音質の自然さ

図 7.1. 聴取実験結果. ＋： 単純加算, A：提案法 A(調波構造に立脚した提案法), B：提案法 B(フォルマント帯域のみを強調する提案法), C：提案法 C(フォルマントに加えて歯擦音も強調する提案法). (c) 一文字違い部での正解率では, 子音の分類別 (All：全て, [s z]：歯擦音, [k g t b]：撥音, [a n m r]：母音および無声音) での結果を併記した

表 7.6. 聴取実験結果へのウィルコクソンの順位和検定の両側検定における p 値. + : 単純加算, A : 提案法 A(調波構造に立脚した提案法), B : 提案法 B(フォルマント帯域のみを強調する提案法), C : 提案法 C(フォルマントに加えて歯擦音も強調する提案法). 片側検定の有意水準 5%として有意な差が認められる数値を太字とし下線を付した

(a) 音節単位での正解率

method	+	A	B	C
+		**0.000**	**0.000**	**0.000**
A			0.336	**0.084**
B				**0.007**
C				

(b) 一文字単位での正解率

method	+	A	B	C
+		**0.000**	**0.000**	**0.000**
A			0.490	0.199
B				**0.061**
C				

(c) 一文字違い部での正解率

method	+	A	B	C
+		**0.000**	**0.000**	**0.000**
A			1.000	0.348
B				0.225
C				

(d) Q1 : 入力音声と比較して, 内容の聴きとりやすさ

method	+	A	B	C
+		**0.000**	**0.000**	**0.000**
A			0.111	0.884
B				**0.037**
C				

(e) Q2 : 入力音声と比較して, 音質の自然さ

method	+	A	B	C
+		**0.000**	**0.000**	**0.000**
A			0.677	0.262
B				0.471
C				

(f) Q3 : 入力 BGM と比較して, 音質の自然さ

method	+	A	B	C
+		**0.000**	**0.000**	**0.000**
A			**0.000**	**0.000**
B				0.576
C				

処理対象とするのに対し, 後者ではフォルマント周波数近傍および歯擦音相応の成分に制限しているためである. また, 提案法 A と提案法 B および提案法 C とでは, 時間周波数平面の同一時間フレームにおける周波数ビン間に設定するゲインの値の緻密さに違いがある. 前者では角周波数ビンについて個別に算出するのに対し, 後者ではフォルマント周波数近傍と歯擦音相応の成分の 2 つのみである.

一方, 聴取実験の集計結果では, 提案法 A と提案法 B および提案法 C とでは, Q3 について前者が劣った以外は, 同等の評価であった. これより, 提案法 A では BGM への処理について冗長であると言える.

入力音声の聴き取りの確保と入力音声および BGM の音質維持の両立を達成するための知見として, 前述の処理内容の違いから, 以下の 2 事項が示唆された. 第一に, 時間周波数平面上での緻密な処理対象の選択こそが重要である. 第二に, 第一の事項が満たされている時, 同一時間フレームにおける周波数ビン間での個別のゲイン係数の算出は, 不利に働く場合がある.

7.3　三手法の特徴のまとめ

実行時間と主観評価実験の結果を, 表 7.7 にまとめる.

表 7.7. 本論文で提案した三手法の特徴

	提案法 A	提案法 B	提案法 C
計算量	◯	×	×
了解度	◯	◯	◎
音声の聴きとりやすさ	◯	△	◯
音声の音質	◯	◯	◯
BGM の音質	△	◯	◯

音声の聴きとりと音質については, 提案法 C が最良である. 一方, 提案法 C は実行時間が入力信号の再生時間の 2 倍以上かかる点が問題である. 実行時間については提案法 A が最良で, 再生時間の 1/6 倍の時間で済む. その上, 提案法 A は BGM の音質以外の評価が提案法 C と同等である.

ここで, 提案法 B のメリットについて考える. まず, 実行時間と主観評価実験について今回検討した観点では, 提案法 B は提案法 A と提案法 C に劣る. 劣った原因は, フォルマント判定部の LPC 係数の計算量が大きかった点にある. 従って, 提案法 B が有利となる条件は, 音声符号化形式などの既に LPC 係数が算出されている信号形態を入力信号として用いた時である. 加えて, 提案法 C は提案法 B に高周波数帯域への処理を追加したものであることから, 高周波数帯域が再生帯域に含まれない音響システムでは, 提案法 B で十分である.

第 8 章
音信号混合法における時間周波数平面の有効性

本稿では, 音のミキシングに時間周波数平面を用いることの有効性を示す. 著者らは, 時間周波数平面を用いた音信号混合法としてスマートミキサーを提案してきた. スマートミキサーは, ミキサー自体が時間周波数平面を用いた解析部を持つことで, 各入力信号の特徴を捉え, 信号間の特徴の相互関係に沿った緻密な処理が実現できる. この考えに立脚したミキシング手法として, 音声と BGM(Background Music) とのミキシング時に, 音声を埋もれさせない音信号混合法を提案してきた. 本章では, スマートミキサーが時間周波数領域での処理であることの有効性を, 提案法 A への聴取実験により検証する.

第 4 章で述べた提案法 A について, 時間周波数領域での処理の有効性を検討する. 提案法 A が生成する時間周波数領域での変化量を, 周波数領域, 時間領域での変化量で近似する方法を定義する. これらに単純加算と提案法 A を加えた 4 種類の混合法について, 聴取実験による了解度と主観評価で比較することで, 時間領域, 周波数領域単独の処理では不十分であることを示す.

8.1 処理量の観点での調波構造に着目した手法 (提案法 A) の説明

ここで, 提案法 A を第 4 章とは異なる切り口で説明する.

処理方法の概要を図 8.1 にまとめる. まず, 入力音声, 入力 BGM, 出力, それぞれの複素時間周波数表現を $X_{\mathrm{A}}[i,k]$, $X_{\mathrm{B}}[i,k]$, $Y[i,k]$ とする. ここで, i は時間フレーム番号,

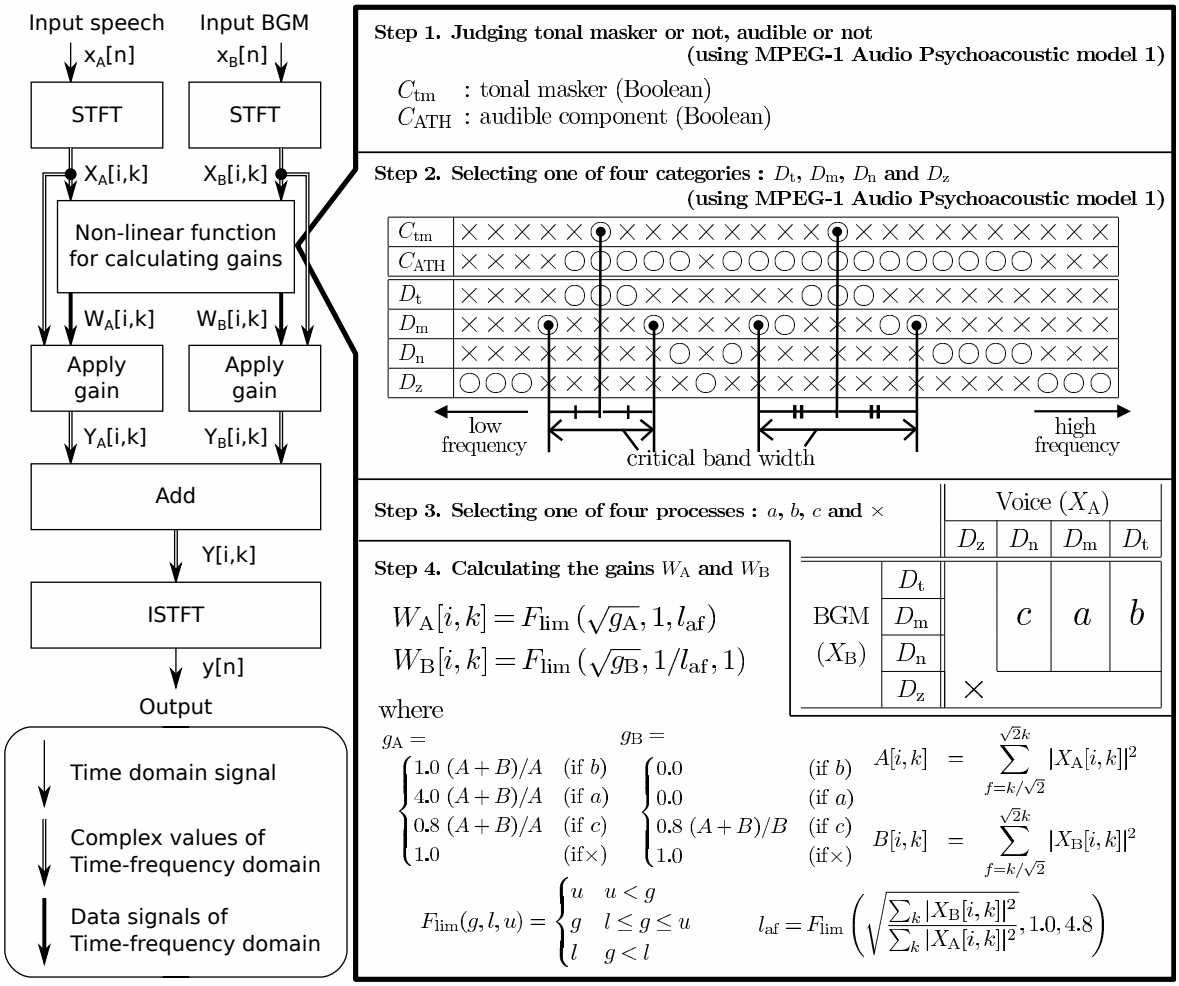


図 8.1. 提案法 A のブロック図およびゲイン生成部における処理の概要

k は周波数ビン番号である. 本手法では, ゲイン係数 $W_{\mathrm{A}}[i,k]$, $W_{\mathrm{B}}[i,k]$ を用いて出力 Y を $Y = W_{\mathrm{A}}X_{\mathrm{A}} + W_{\mathrm{B}}X_{\mathrm{B}}$ で生成する.

ゲイン係数を決めるにあたって, Step 2 では, 各 $[i,k]$ について $X_{\mathrm{A}}[i,k]$, $X_{\mathrm{B}}[i,k]$ を以下の 4 つのカテゴリー, 純音 D_{t}, 純音近傍 D_{m}, 非純音 D_{n}, 非可聴 D_{z} に分類する. その分類のために, Step 1 では, MPEG-1 Audio Psychoacoustic model 1 [59] で採用されている純音判定 C_{tm} と有音判定 C_{ATH} の基準を利用している.

Step 3 では, 音声と BGM の 4 つのカテゴリーの組み合わせで決まる $16(=4\times4)$ 属性のそれぞれに対して対応表 (図 1 の Step 3 欄を参照) を使って, 4 つの処理 a, b, c, $\times$ を対応させる.

これら 4 つの処理について順に説明する. まず, $\times$ では少なくとも片方の入力が無音であるから単純加算で良い. 次に, a, b, c に対するゲイン係数設定法について述べる準備として, X_{A}, X_{B} の $[i,k]$ におけるパワーを周波数方向に平滑化したものをそれぞれ $A[i,k]$, $B[i,k]$ とおく. 出力パワーは, 互いの位相がランダムであるとすれば $P = W_{\mathrm{A}}^2A + W_{\mathrm{B}}^2B$ と見積もることができる. また, 出力パワー中の X_{A} 成分の割合は $Q = W_{\mathrm{A}}^2A/P$ で見積

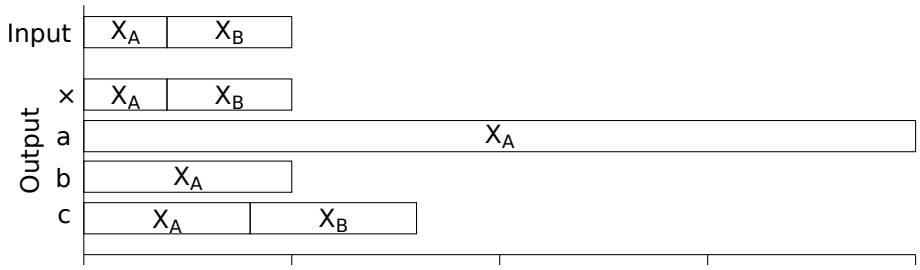


図 8.2. P, Q の決定法. X_{A}, X_{B} は, 由来する入力信号を示す

表 8.1. 実験諸元

Sampling frequency (F_{S})	44100 Hz
Samples of FFT (N_{FFT})	1024
Samples of frame shift	384
Type of analyze window	Hanning
Samples of analyze window	1023
Type of synthesize window	Hanning
Samples of synthesize window	767

もることができる. 我々は, P, Q を A, B の関数として適切に決定することで, 音声が埋もれない混合ができるものと考えた.

P, Q は図 8.2 に従って決定する. まず, a では $P = A + B$, $Q = 1.0$ とし, 出力を強制的に音声化する. 次に, b では $P = 4.0(A + B)$, $Q = 1.0$ として, a に比較して P を 4.0 倍することで混合音においてもピークを保存することを狙った. 一方, c では音声は重要ではないため, $P = 1.6(A + B)$, $Q = 0.5$ とし, 音声と BGM を等音量化するに留める. なお, 係数である 4.0 と 1.6 は予備実験により決定した.

Step 4 では, 以上で求めた P, Q から W_{A}, W_{B} を逆算する. ただし, 音声の抑制および BGM の強調が行われないよう, $W_{\mathrm{A}} \geqq 1.0$, $W_{\mathrm{B}} \leqq 1.0$ の条件をつける. また, 過度の音声の強調および BGM の抑制が起こることを防止するいくつかの条件を設けているがここでは説明を省く. 提案法 A の諸元を表 8.1 に示す.

8.2　時間領域, 周波数領域での近似処理

本章では, 提案法 A が時間周波数領域でゲイン調整を行っていることの利点を示すために, それを周波数領域に限定して近似したゲイン調整 (以下, イコライザ近似と言う) と, 時間領域に限定して近似したゲイン調整 (以下, ダッカー近似と言う) を定義する.

まず, 準備として $Y_{\mathrm{A}}[i,k]$, $Y_{\mathrm{B}}[i,k]$ を, それぞれ ISTFT により時間信号 $y_{\mathrm{A}}[n]$, $y_{\mathrm{B}}[n]$ に

変換する. 次に, $y_{\mathrm{A}}[n]$, $y_{\mathrm{B}}[n]$ を, 提案法 A と同一諸元 (表 8.1) の STFT で再解析し, その結果をそれぞれ $V_{\mathrm{A}}[i,k]$, $V_{\mathrm{B}}[i,k]$ とする.

イコライザ近似では, まず, 時間周波数平面の観察により音声と BGM がともに有音であると判断したフレームの集合 $\mathbb{I}^{\mathrm{E}}$ について, 周波数ビンごとのパワーの総和が提案法 A に等しくなるゲイン $G_{\mathrm{c}}^{\mathrm{E}}[k]$ を以下によって求めておく.

$$G_{\mathrm{c}}^{\mathrm{E}}[k] \;=\; \sqrt{\frac{\frac{1}{I^{\mathrm{E}}}\sum_{i\in\mathbb{I}^{\mathrm{E}}}|V_{\mathrm{c}}[i,k]|^2}{\frac{1}{I^{\mathrm{E}}}\sum_{i\in\mathbb{I}^{\mathrm{E}}}|X_{\mathrm{c}}[i,k]|^2}} \tag{8.1}$$

ここで I^{E} は, $\mathbb{I}^{\mathrm{E}}$ の総フレーム数である. また, c は A または B である. G^{E} を用いて, イコライザ近似での出力信号の時間周波数表現 $Z^{\mathrm{E}}[i,k]$ を,

$$Z^{\mathrm{E}}[i,k] \;=\; G_{\mathrm{A}}^{\mathrm{E}}[k]X_{\mathrm{A}}[i,k] + G_{\mathrm{B}}^{\mathrm{E}}[k]X_{\mathrm{B}}[i,k] \tag{8.2}$$

として算出する.

ダッカー近似では, まず, フレームごとのパワーの総和が提案法 A に等しくなるゲイン $G_{\mathrm{c}}^{\mathrm{D}}[i]$ を以下によって求めておく.

$$G_{\mathrm{c}}^{\mathrm{D}}[i] \;=\; \sqrt{\frac{\frac{1}{N_{\mathrm{FFT}}}\sum_{k=0}^{N_{\mathrm{FFT}}-1}|V_{\mathrm{c}}[i,k]|^2}{\frac{1}{N_{\mathrm{FFT}}}\sum_{k=0}^{N_{\mathrm{FFT}}-1}|X_{\mathrm{c}}[i,k]|^2}} \tag{8.3}$$

G^{D} を用いて, ダッカー近似での出力信号の時間周波数表現 $Z^{\mathrm{D}}[i,k]$ を,

$$Z^{\mathrm{D}}[i,k] \;=\; G_{\mathrm{A}}^{\mathrm{D}}[i]X_{\mathrm{A}}[i,k] + G_{\mathrm{B}}^{\mathrm{D}}[i]X_{\mathrm{B}}[i,k] \tag{8.4}$$

として算出する.

$G_{\mathrm{c}}^{\mathrm{E}}[k]$, $G_{\mathrm{c}}^{\mathrm{D}}[i]$ 算出の概念図を, 図 8.3 に示す.

以上のようにして求めた $Z^{\mathrm{E}}[i,k]$, $Z^{\mathrm{D}}[i,k]$ を ISTFT することで, イコライザ近似とダッカー近似の出力信号を得る.

8.3　聴取実験による近似処理との比較

単純加算 (A), 提案法 A(S), イコライザ近似 (E), ダッカー近似 (D) の 4 手法について, 聴取実験による了解度と主観評価で比較する. 被験者は成人男性 12 人であり, ヘッドホン (DENON AH-D2000) によって音を提示した.

ここで, 比較対象に従来法として自動ミキサーを含めなかった理由について述べる. これは, 予備実験として自動ミキサーでの処理を模擬した混合結果が, 提案法に比べて著しく BGM の音量が小さく, 評価対象に含めるべきでないと判断したからである. 第 2 章での検討から, 自動ミキサーが一般に信号の切り替えを行うものであるか, ダッカーであるとまとめた. 自動ミキサーでのゲインは, 優先度の高い入力信号へのゲイン調性は行われず, 優先度の低い入力信号への抑制のみとするのが主である. 従って, 入力時

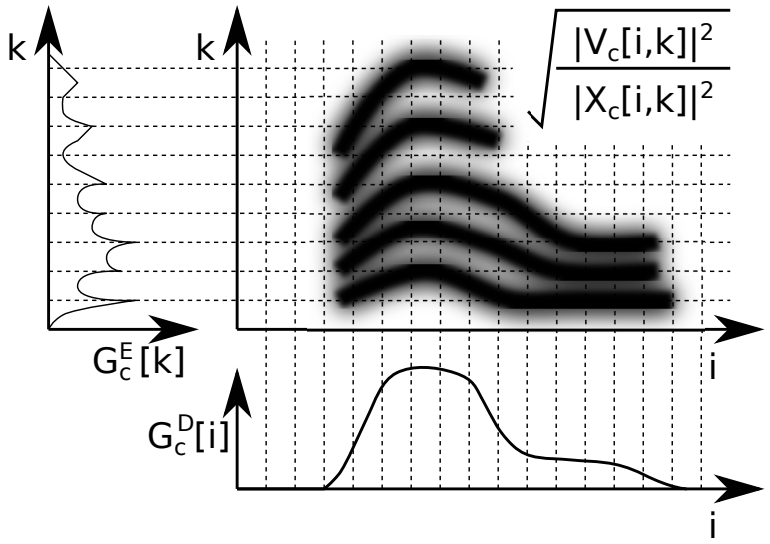


図 8.3. $G_c^{\mathrm{E}}[i]$, $G_c^{\mathrm{D}}[i]$ 算出の概念図

に音声がBGMよりも音量が小さい場合では, BGMは音声よりもさらに小さくなり, 不自然な処理結果となる. 今回実施する実験での音声とBGMの音量差の設定 (詳しい設定は後述する) が, このケースに合致するため, 従来法を比較対象に含めることができなかった.

入力信号は, 表8.2 に示す3セットを用いた. 音声は「話速バリエーション型音声データベース SRV-DB[55] 」の, 「発話のプロフェッショナルによるオリジナル原稿（カーナビ文章）の読み上げ」から, 09 ～ 32 文の 24 データを用いた. BGMは「 RWC 研究用音楽データベース: ポピュラー音楽 [56] 」に収録の音源を用いた. 各セットでの音声とBGMとの音量差は, ある一人の被験者に対して単純加算したときに音声が聴きとれないように予め調整した. また, 音声にSNRが −10 dBの白色雑音を付加した実験も行った. 白色雑音はピンクノイズほど騒音を代表する信号ではないが, 一部の騒音 (例えばフライパンによる調理時の音) に対しては類似性が高い. 音声の 24 データのうち, 09 ～ 20 文の 12 データに白色雑音を付加し, 21 ～ 32 文の 12 データには白色雑音を付加せず用いた.

サンプルは実験計画法により, 各被験者が聴取する文の回数が 1 回ずつ重複せず, BGM3 種類とミキシング手法 4 種類との組み合わせ数が等しくなるように振り分けた. 被験者には, サンプルを 1 度だけ聴かせ, 聴きとれた内容をひらがなで記入させた. 了解度の算出は, サンプルごとに 3 段階 (2:全音節正解, 1:一部音節正解, 0:全音節不正解) で得点づけし, 全サンプルで全音節正解の場合に, 100%となるよう正規化した.

主観評価は, Q1 : “入力音声と比較して, 内容の聴きとりやすさ”, Q2 : “入力音声と比較して, 音質の自然さ”, Q3 : “入力BGMと比較して, 音質の自然さ” の 3 つの評価項目とした. 評価値は, 表 8.4 に示す 5 段階の DMOS(Degradation Mean Opinion Score) 値 [67] とした.

入力信号は, 表8.3 に示す3セットを用いた. 音声には「話速バリエーション型音声

表 8.2. 了解度での, 入力信号 3 セットの設定

Set		Contents	Standard deviation	Volume difference
I1	Voice	PF02 (Female)	500	18 dB
	BGM	No.5	4000	
I2	Voice	PF02 (Female)	375	21 dB
	BGM	No.15	4000	
I3	Voice	PF02 (Female)	375	24 dB
	BGM	No.94	6000	

表 8.3. 主観評価での, 入力信号 3 セットの設定

Set		Contents	Standard deviation	Volume difference
E1	Voice	PF00 (Female)	500	12 dB
	BGM	No.5	2000	
E2	Voice	PM00 (Male)	1000	12 dB
	BGM	No.15	4000	
E3	Voice	PF01 (Female)	375	24 dB
	BGM	No.94	6000	

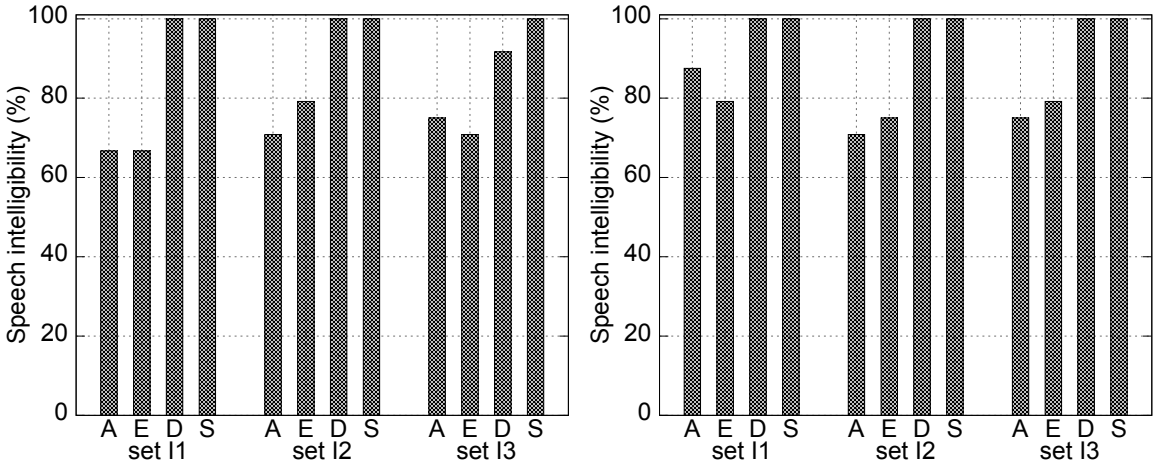
データベース SRV-DB[55] 」の「発話のプロフェッショナルによる編集手帳（読売新聞）の読み上げ」を用いた. 4.1 節の実験と異なる音声を使用したのは, カーナビ音声は文が短く抑揚が単調で, 音質の自然さの評価が困難であるためである. 一方, BGM には 4.1 節の実験と同じものを用いた. 各セットでの音声と BGM との音量差は, ある一人の被験者に対して単純加算したときに音声が聴きとれないように予め調整した. さらに入力信号の音声について, 了解度と同じく SNR が -10 dB の白色雑音を付加した音源もあわせて用意した. 全サンプルを全被験者に聴取させた.

了解度の結果を図 8.4 に示す.

まずイコライザ近似では, 単純加算からの改善が見られない. 提案法 A とダッカー近似で, 白色雑音の付加の有無にかかわらず 100%を達成した. 提案法 A が了解度を向上させている要因として, 時変なゲイン操作が大きな役割を果たしていることが示された.

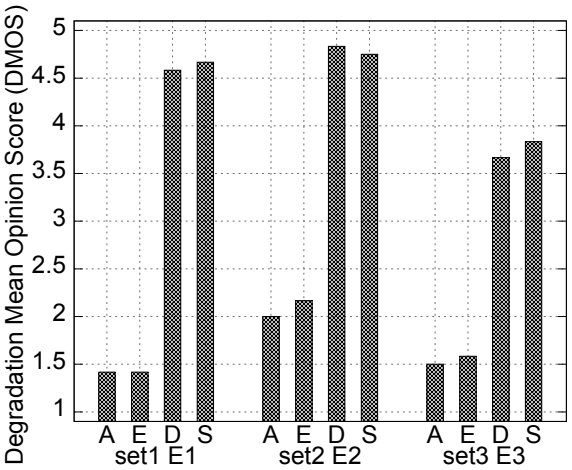
表 8.4. 主観評価での, DMOS 値の設定

Point	Impairment
5	Degradation is inaudible
4	Degradation is audible but not annoying
3	Degradation is slightly annoying
2	Degradation is annoying
1	Degradation is very annoying

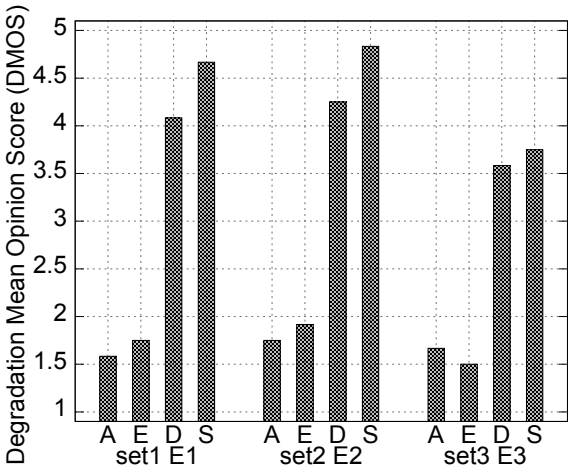


(a) Without noise　(b) With noise

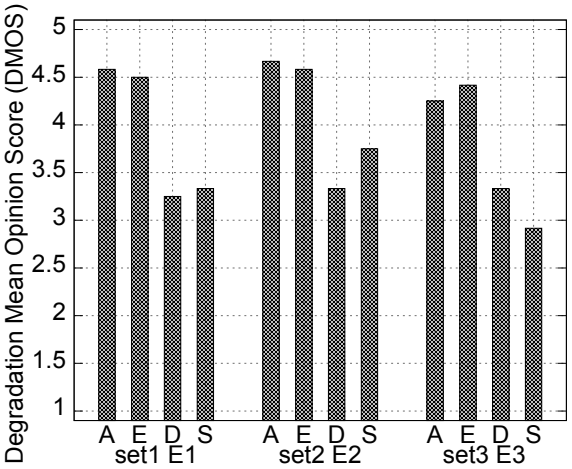
図 8.4. 了解度の聴取実験結果. 左 (a) : 白色雑音付加なし, 右 (b) : 白色雑音付加あり. A : 単純加算, E : イコライザ近似, D : ダッカー近似, S : 提案法 A



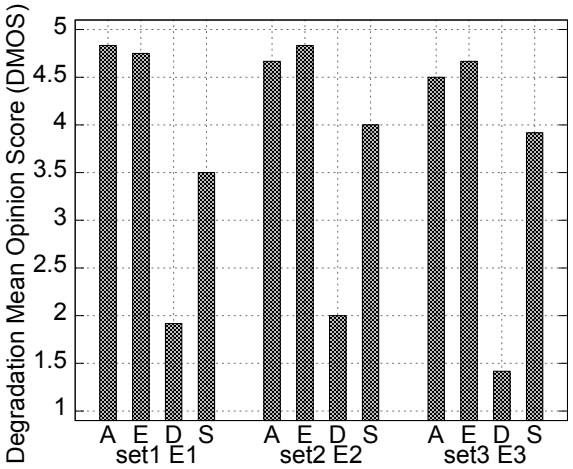
(a) Q1 : Voice audibility, Without noise



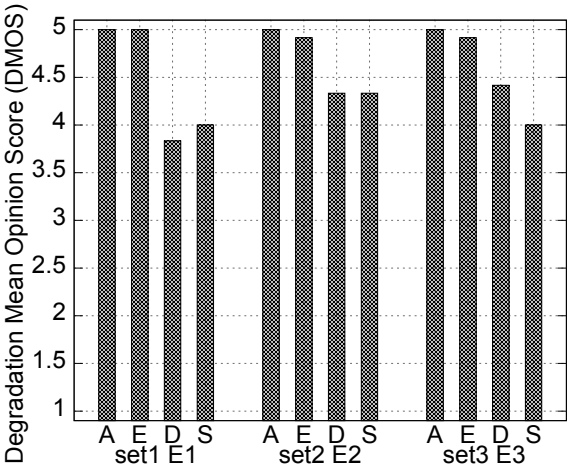
(b) Q1 : Voice audibility, With noise



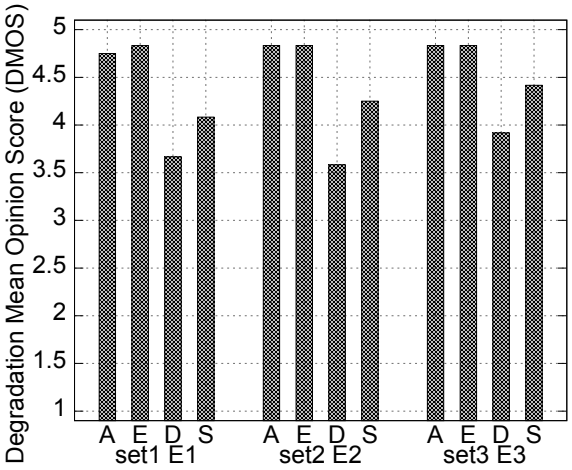
(c) Q2 : Sound quality of the voice, Without noise



(d) Q2 : Sound quality of the voice, With noise



(e) Q3 : Sound quality of the BGM, Without noise



(f) Q3 : Sound quality of the BGM, With noise

図 8.5. 主観評価の聴取実験結果. 上 (a,b) : Q1 音声の内容の聴きとりやすさ, 中 (c,d) : Q2 音声の音質の自然さ, 下 (e,f) : Q3 BGM の音質の自然さ. 左 (a,c,e) : 白色雑音付加なし, 右 (b,d,f) : 白色雑音付加あり. A : 単純加算, E : イコライザ近似, D : ダッカー近似, S : 提案法 A

主観評価の結果を図 8.5 に示す.

“音声の内容の聴きとりやすさ (Q1)” については, (a) より, 提案法 A とダッカー近似が秀でている. 一方, 白色雑音付加ありでの評価 (b) を (a) と比較すると, ダッカー近似でのみ全セットで評価が下がっている.

“音声の音質の自然さ (Q2)” については, (c) より, 単純加算とイコライザ近似では評価が高く, ダッカー近似と提案法 A では低い. 一方, 白色雑音付加ありでの評価 (d) を (c) と比較すると, ダッカー近似のみ評価が下がり, 提案法 A では逆に評価が上がっている.

“BGM の音質の自然さ (Q3)” については, (e) より, 単純加算とイコライザ近似では評価が高く, ダッカー近似と提案法 A では低い. 一方, 白色雑音付加ありでの評価 (f) を (e) と比較すると, 提案法 A のみ評価が上がり, ダッカー近似では逆に評価が下がっている.

Q2 と Q3 の比較から, 音声への白色雑音の付加による音質の自然さの劣化は, BGM の劣化としても評価されたことがわかる. 提案法 A では, 音声, BGM ともに評価が向上している. 逆にダッカー近似では, 音声, BGM ともに評価が低下している.

以上から, 時間領域でのゲイン操作により, 音声の了解度と内容の聴きとりやすさを得られることが示された. 周波数帯域ごとに処理を異にすることで, 音質劣化を低減できることが示された. そして, 両者を組み合わせた時間周波数平面での処理により, 聴きとりやすさと音質の両立が実現できることが示された.

8.4　聴取実験結果を受けての考察

聴取実験の結果から得た知見を, 時間周波数平面上での観察と照らし合わせる. 入力信号として表 8.3 の E3 を用いたときの入出力信号の時間周波数平面を図 8.6 に示す (パワーが大きい部分を黒で表している). 図 8.6 の 12 枚の時間周波数平面のうち, 左の 6 枚は白色雑音を加えない場合の入出力信号であり, 上から入力音声 (CV), 入力 BGM(CB), 単純加算 (CA), イコライザ近似 (CE), ダッカー近似 (CD), 提案法 A(CS) である. 右の 6 枚は白色雑音を加えた場合の入出力信号であり, 上から入力音声 (WV), 入力 BGM(WB, ただし CB と同一である), 単純加算 (WA), イコライザ近似 (WE), ダッカー近似 (WD), 提案法 A(WS) である. これら 12 枚の図に対して考察の対象とする 4 領域 α, β, γ, δ を破線で囲った. α には音声の母音認識で重要なフォルマントと周期成分, β には音声の話者性に関連するフォルマントと周期成分, γ には音声の非周期成分, δ には BGM のアタック成分が存在する. 各領域内での濃淡の模様に注目し, 入力信号の濃淡の模様が, 出力信号の濃淡の模様として明瞭に視認できれば, その成分は出力信号でも聴きとりやすいと考える.

あわせて, 信号間での模様の比較を, 数値指標でも確認する. 第一の数値指標として, 領域内の各点でのパワー (dB 表現) の相関係数を用いた (表 8.5). 第二の数値指標として, 領域内での平均パワーを用いた (表 8.6).

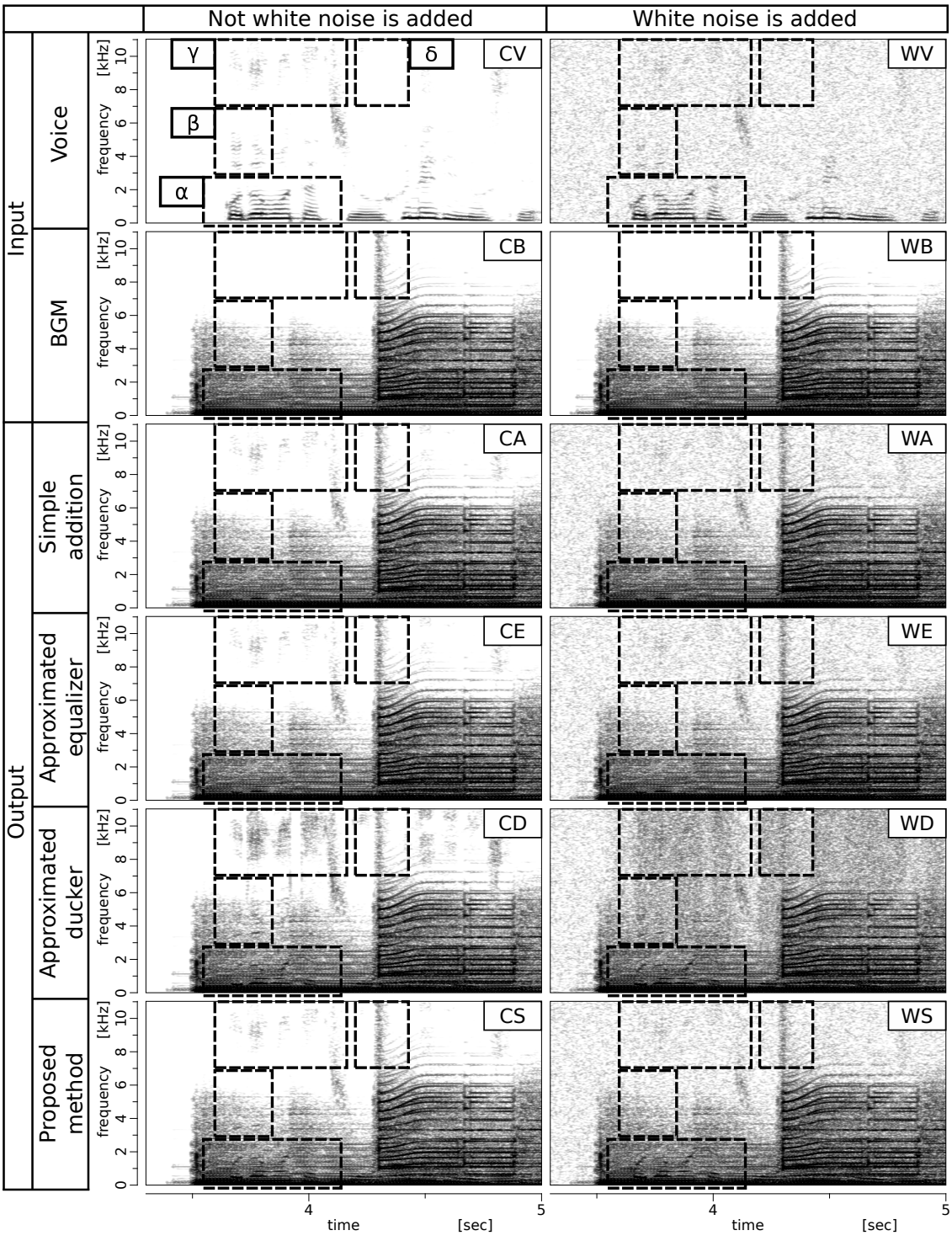


図 8.6. 入出力信号の時間周波数平面. 右：白色雑音付加なし, 左：白色雑音付加あり. 上から, 入力音声, 入力 BGM, 単純加算, イコライザ近似, ダッカー近似, 提案法 A

表 8.5. 図 8.6 の枠内における入出力信号間についての相関係数

(a) α , Without noise

	CV	CB
CA	0.226	0.965
CE	0.227	0.964
CD	0.347	0.881
CS	0.325	0.855

(b) α , With noise

	WV	WB
WA	0.237	0.952
WE	0.239	0.951
WD	0.415	0.830
WS	0.439	0.787

(c) β , Without noise

	CV	CB
CA	0.467	0.948
CE	0.467	0.948
CD	0.612	0.822
CS	0.491	0.929

(d) β , With noise

	WV	WB
WA	0.420	0.648
WE	0.420	0.648
WD	0.733	0.236
WS	0.473	0.631

(e) γ , Without noise

	CV	CB
CA	0.955	-0.150
CE	0.955	-0.150
CD	0.912	-0.193
CS	0.955	-0.151

(f) γ , With noise

	WV	WB
WA	0.997	-0.038
WE	0.997	-0.039
WD	0.860	-0.097
WS	0.997	-0.039

(g) δ , Without noise

	CV	CB
CA	0.049	0.929
CE	0.048	0.929
CD	0.241	0.781
CS	0.049	0.928

(h) δ , With noise

	WV	WB
WA	0.403	0.581
WE	0.403	0.581
WD	0.749	0.106
WS	0.413	0.583

まず, α と β に着目する. 表 8.6(a), (c) より, 入力 CV に対する出力 {CA, CE, CD, CS} の相関は, CA の場合に比較して, CE では改善が見られず CD と CS では改善が見て取れる. これは, 音声の内容の聴きとりやすさの結果 (図 8.5(a)) と一致している. す

表 8.6. 図 8.6 の枠内における入出力信号の平均パワー

(a) Without noise (dB)

	α	β	γ	δ
CV	58.4	33.2	31.8	17.5
CB	82.7	43.9	5.9	42.7
CA	82.8	44.3	31.8	42.7
CE	82.4	44.3	31.8	42.7
CD	82.4	46.2	42.4	41.7
CS	82.3	44.6	31.8	42.7

(b) With noise (dB)

	α	β	γ	δ
WV	58.4	38.6	38.1	37.4
WB	82.7	43.9	5.9	42.7
WA	82.8	45.1	38.1	43.8
WE	82.3	45.1	38.1	43.8
WD	82.1	50.2	48.4	48.9
WS	82.0	46.7	38.1	43.8

なわち, イコライザ近似に対してのダッカー近似と提案法 A の優位性が数値指標によって裏付けられた.

次に, γ に着目する. 図 8.6 を見れば明らかに WD でのみ γ でのパワーが突出している. 実際, 平均パワー (表 8.6(b)) は WA に比べて約 10 dB($\approx 48.4 - 38.1$) 増えている. 一方, 表 8.5(f) より WV との相関が WD でのみ下がっていることから, その音声は劣化している. これらの結果は, 音声の音質の自然さ (図 8.4(d)) での結果と一致し, ダッカー近似に対しての提案法 A の優位性が数値指標によって裏付けられたことになる.

続いて, δ に着目する. 図 8.6 より BGM のアタック成分は WD でのみ消失している. 表 8.5(h) より WB と WD の相関が 0.106 になっていることからも, BGM が劣化していると確かめられた. これは, BGM の音質の自然さ (図 8.4(f)) での結果と一致し, ここでもダッカー近似に対する提案法 A の優位性が示されたことになる.

時間周波数平面での観察により, 以下のことがわかった. まず, 音声の聴きとりやすさの確保では, 出力信号での音声の時間周波数平面での濃淡の模様の維持が重要であると考え, それは主観評価の実験結果と一致した. 一方, 音質の評価の維持では, 背景雑音を処理の適用外とすることが重要である. 提案法 A では両者を達成できており, 提案法 A のイコライザ近似, ダッカー近似では達成できない. よって, 提案法 A での時間周波数領域での処理の有効性が示された.

なお, イコライザ近似, ダッカー近似での変化量は, 提案法 A での時間周波数領域での処理に基づいて算出している. 従って, ダッカー近似が提案法 A と同等の評価になる場合があったとしても, 計算量の観点でダッカー近似法は実用的ではない.

8.5　時間周波数平面の有効性についての結論

本章で, 音声と BGM のミキシングについて音声を埋もれさせない音信号混合法を実現する方法として, 時間周波数領域でのゲイン調整が有効であることを示した.

音声と BGM のミキシングを, 単純加算, 提案法 A, イコライザ近似, ダッカー近似の 4 手法で実施し, 聴取実験により了解度と主観評価を比較した. 音声は, 白色雑音を付加した音源とそうでない音源との 2 種類で検討した.

白色雑音付加なしでは, 提案法 A とダッカー近似が同等の良い評価を得た. 一方, 白色雑音付加ありでは, ダッカー近似では評価が低下した. 時間周波数領域でのゲイン調整により, 音声に混入する背景雑音レベルに依らず, 良好な結果が得られることが示された.

また, 時間周波数平面での観察と客観的な数値指標によって, 聴取実験での評価を裏付けることができた. 以上の実験により, 音声の聴きとりやすさと音質の両方が, 提案法 A によってのみ達成できることを明らかにした. すなわち, 時間周波数領域での処理の有効性が示された.

第 9 章
結論と今後の課題

9.1 結論

本研究の目的は, 第一に, 音声を埋もれさせない音信号混合を実現することであり, 第二に, 音信号混合に時間周波数平面を用いることの有効性を示すことである. ここで, 音声を埋もれさせない音信号混合の定義を, 音声と BGM(Background Music) を入力した時に, 第一に音声の内容の聞き取りを維持するものとし, 第二に音声と BGM の音質を維持するものとする. 本論文では, これらの目的を達成するために, 時間周波数平面を用いた音信号混合法を 3 手法提案し, 主観評価実験と時間周波数平面の観察により有効性を検討した.

第 2 章で, 音楽制作におけるミキシング処理についてまとめた. まず, ミキシング処理の目的が, 音量, 定位, 音色のバランスを整え, 主役をはっきりさせることにあるとまとめた. 続いて, 従来のミキシング処理として 5 例を例示し, 従来のミキシング処理が時間領域処理と周波数領域処理の組み合わせによる時間周波数領域処理にとどまっていることを示した. その問題点として, 時間領域処理と周波数領域処理の組み合わせでは実現できない時間周波数領域処理がある点を指摘した.

第 3 章で, スマートミキサーについて述べた. スマートミキサーの定義は, 時間周波数平面を用いた解析部をミキサー内部に持ち, ミキシングを時間周波数平面上での重ね合わせとして実施する音信号のミキサーである. スマートミキサーと従来ミキサーとを比較し, スマートミキサーとしての利点が, エフェクト処理の相互作用を周波数帯域ごとに勘案できる点にあるとまとめた. また, スマートミキサーの実装について, ハードウェアとソフトウェアについて例示した.

第 4 章で, 調波構造に着目した手法を提案法 A として提案した. この手法では, 時間

周波数平面上での調波構造のパターンを捉えるために, MPEG1 Audio に含まれる聴覚心理モデルの純音判定を基準とした. この提案法 A について, 出力信号の時間周波数平面の観察と聴取実験による主観評価実験を行った. 評価実験では, 提案法 A の時間周波数平面上での処理量を時不変のイコライザとして近似する処理法を比較対象とした. 比較により, 音声の聴き取りやすさについて, 提案法 A の評価が上回る (ウィルコクソンの順位和検定の片側検定において, $p = 0.027$) ことがわかった.

第 5 章で, フォルマント構造に着目した手法を提案法 B として提案した. この手法では, 音声の母音の識別性を担うフォルマント構造を基準とした. フォルマント構造は調波構造と比較して時間周波数平面上での領域が狭いことから, 聴きとりと音質を両立できると考えた. 主観評価実験によって有効性を確かめるとともに, 子音のうち特に歯擦音に聴き取りづらさが残ると考察した.

第 6 章で, フォルマント構造と歯擦音に着目した手法を提案法 C として提案した. この手法は, 提案法 B によるフォルマント構造への処理に, 歯擦音への処理を追加したものである. 歯擦音への処理は, 母音の周波数分布の特徴は 3 kHz より低い周波数帯域で顕著であり, 歯擦音の周波数分布の特徴は 4 kHz よりも高い周波数帯域で顕著であることから, 4 kHz より高い周波数帯域において歯擦音のスペクトル遷移パターンを強調するものとした. 同一女性話者による子音のみが異なる 3 種の音声信号を用いた実験により, 提案法 C が適切に時間周波数平面上での歯擦音のパターンを強調できることを示した.

第 7 章では, 提案した 3 手法について比較と考察を行った. 第一の比較として, 実行時間を計測した. 第二の比較として, 主観評価と発話内容の了解度の聴取実験を行った. 提案した 3 手法が, 音声の聴き取りやすさ, 音声の音質の自然さ, BGM の音質の自然さ, 了解度の全ての項目で, 単純加算よりも良い評価 (ウィルコクソンの順位和検定の両側検定において, 全項目で $p = 0.000$) を得た. 提案法 A と提案法 C で同等の音声の聴き取りやすさ (ウィルコクソンの順位和検定の両側検定において, $p = 0.884$) が確保できることが示された. 提案法 A と提案法 C の評価は BGM の音質劣化の項目のみで分かれ, 調波構造に着目した手法がより音質劣化が目立つ (ウィルコクソンの順位和検定の両側検定において, $p = 0.000$) と評価された. これらの比較から, 音声の聴きとりやすさと音質については提案法 C が最良であることがわかった. また, 計算量については提案法 A が最小であり, 提案法 C と同等の音声の聴きとりやすさの評価が得られることがわかった. 一方, 再生帯域の上限が低い音響システムで用いる場合や, 音声符号化形式を用いる場合では, 提案法 B が有利である.

第 8 章で, スマートミキサーにおける時間周波数平面利用の有用性を確認した. 提案法 A の時間周波数平面上での処理量を周波数軸に投影した手法 (イコライザ近似), 時間軸に投影した手法 (ダッカー近似) を定義し, 時間周波数平面での観察と客観的な数値指標で比較した. さらに, 入力音声信号には SNR が -10 dB となる白色雑音を加算したものも用意し, 背景雑音レベルの高い環境下での集音の模擬を狙った. 比較の結果, 音声の聴きとりやすさと音質の維持が提案法 A によってのみ達成できることを明らかにし, 時間周波数領域での処理が有効であることを示した.

以上より, 本論文の提案法は音声を埋もれさせない音信号混合を実現できることが示された. また, 音信号混合において時間周波数平面の利用の有効性が, 本論文の提案法における時間周波数平面上での処理領域制限の効果として示された.

9.2　今後の課題

今後の課題として第一に, 本論文で主として提案した 2 手法の知見を総合した手法の検討が挙げられる. 第 7, 8 章での考察から, 時間周波数平面での処理領域制限の重要性を確認した. BGM の音質の項目で, 調波構造に着目した手法が劣った. 調波構造に着目した手法とフォルマントに着目した手法とでは, 共通する処理領域における処理量が異なる. ここまでの検討で, 調波構造への着目が不要であるとまでは結論付けられない. 両手法の違いを整理し, 一つずつ有効性を検討する必要がある.

第二に, フォルマント周波数推定法の代替法の検討が挙げられる. 本論文で述べた手法では LPC を用いており, 計算量が大きい. LPC は音声の生成モデルと対応が良いため, 代替法を用いることでフォルマント推定の劣化する. フォルマント推定の劣化に伴って, 音質も聞き取りやすさも低下するだろう. 従って, フォルマント推定での誤差の許容量と音質, 計算量の関係を調べる必要がある.

第三に, 処理領域制限における時間周波数平面の分解能の検討である. 本論文の提案法での時間周波数平面の諸元は, 聴覚心理モデルの仕様に因る. 処理結果への主観評価, 計算量の観点に立った検討が実施できる.

第四に, 適用可能な入力信号の種類の拡張である. 本論文の提案法では, 音声を埋もれさせない音信号混合法の実現を主眼としていた. このため, 提案法で着目する特徴として音声との関係があるものを選別した. その上, 評価実験においても入力信号を音声と BGM の組み合わせのみを対象とした. 異なる特徴に基づいた手法の検討は元より, 本論文の提案法についても検討の余地がある. 例えば, 調波構造に基づいた手法では, 調波構造が優勢な音であれば対応可能である.

第五に, 多様な音響空間への対応である. 本論文では全てのパラメータ調整, 聴取実験をヘッドホンによって行った. しかし, ヘッドホンとスピーカとでは適切なパラメータに違いがあり, さらに再生する部屋によっても異なることがわかっている. 実用化に向けて, 利用する音響空間に合わせた処理を実現する必要がある.

第六に, リアルタイム実装である. 実用化はもちろんのこと, 提案法を効率良く改良するためにも, リアルタイム実装が重要である. リアルタイム実装ではレイテンシは可能な限り短いのが良い. FPGA を用いたハードウェア上でのリアルタイム実装の検討により, レイテンシが解析窓幅の 1/2 程度で実装できることがわかっている. このレイテンシは, 処理対象とするサンプルを中心とした時間周波数表現を得るために必要不可欠である. 従って, レイテンシを所望の値とするためには, レイテンシに対応する解析窓幅での時間周波数表現に合わせた処理が必要である.

謝辞

指導教員の高橋弘太准教授に深く感謝いたします. 本研究を行うきっかけを与えて下さり, 研究の場を提供して下さいました. 漲る好奇心と情熱をもって, 活発な議論, 手厚い指導を授けて下さいました. 指導のみならず研究生活を進める上での体調管理, 精神衛生についても, 惜しみなくご助言くださいました.

本論文の作成にあたり, 数多の有益なご意見をくださいました三橋渉名誉教授, 張熙教授, 肖鳳超教授, 野村英之准教授に心よりお礼申し上げます.

高橋弘太研究室の皆様に感謝いたします. 学部 3 年から 7 年間に渡り, 才能も個性も豊かな方々と議論する機会に恵まれ, 多くのことを学ばせていただきました. 有北知弘氏と宮地紘司氏に感謝いたします. ご自身の研究や活動がお忙しい最中にも, 私の学外発表の時期などでは, 大いに支えて下さいました. 中でも, 2012 年 11 月のレジュメ締め切り間際に戴いたプリンのことは忘れられません. 手塚歩氏と旭岡舜介氏に感謝いたします. ミキシングエンジニアとして, 実験サンプルの制作にもご協力いただきました.

元電気通信大学知的財産部門の尾原和貴様, 元 JST の前田武志様をはじめ, スマートミキサーの知的財産権化にご尽力くださいました皆様に感謝いたします. 打ち合わせを通して, スマートミキサーへの理解が深まり, 新規性の整理ができました.

電気通信大学産学官連携センターの今田智勝様, キャンパスクリエイトの益田忍様をはじめ, スマートミキサーの産学連携に御尽力くださった皆様に感謝いたします. スマートミキサーに篤く共感いただき, 多くの応用の可能性を示していただき, 大変勇気づけられました.

西方夏子様に感謝いたします. iPhone アプリ omoide として, スマートミキサーを初めて実用化してくださいました. 実用化に耐え得るように, 想定すべきシチュエーションや音質について, 数多くの有益なご意見をいただきました. また, 濃密な議論の時間

を通じて, 大いに刺激を受けました.

葛巻善郎氏と宮地楽器神田店の方々に感謝いたします. 音楽制作のミキシングの技を披露してくださるだけでなく, 本研究のデモについても貴重なご助言をくださいました.

電子情報通信学会応用音響研究会, 日本音響学会電気音響研究委員会, 日本音響学会騒音・振動研究委員会の方々に感謝いたします. 研究会での発表の折には, 温かいご指導をいただきました.

学外研究会での発表や研究室公開で, 興味を持って説明を聴いてくださった方々に感謝いたします. 大変励みになりました.

電気通信大学ものつくりセンター電子回路設計工作部門の梶川竜義様, 青木猛様, 電子工学工房受講生の方々に感謝いたします. 電子工作に興味がありつつも, 個人では二の足を踏みがちでした. 悶々としていた私にとって, 工作室で日々見かける作品や工具は, 大変刺激的で勉強になりました.

学科事務の方々にも感謝いたします. 研究室や TA での用事で, いつもお世話になりました.

大学生協にも感謝いたします. 音響系の工学書の新刊を, いつも入荷してくださり, 大変助かりました. 特に MPEG の規格書については, これまで購入事例がなかったにもかかわらず, 取り扱ってくださいました. MPEG の規格書なくしては, 本論文はありませんでした.

白井洋佐氏に感謝いたします. ギター制作や音楽制作にと, まわり道に勤しんでいた当時の私に, 高橋先生が担当されていた信号処理論の講義の内容と評判を, 教えて下さいました. その情報こそが, 私が信号処理に興味を持ったきっかけでした.

森直樹氏に感謝いたします. 私と共に音楽ユニット IED として, 音楽作品の製作に付き合ってくださいました. 異なる音楽趣味を持ち, 私の音楽への視野を拡げてくださいました.

両親に感謝いたします. 院進学に理解を示して下さり, 生活基盤を与えていただけたからこそ, 私は研究に打ち込むことができました.

本研究は, RWC 研究用音楽データベース (ポピュラー音楽) より, RWC-MDB-P-2001-M01 No.5, RWC-MDB-P-2001-M01 No.15, RWC-MDB-P-2001-M07 No.94 を利用しました. 後藤真孝先生をはじめとした, RWC 音楽データベースサブワーキンググループに感謝いたします.

付録A
MPEG-1/Audio Psychoacoustic model 1の抜粋

本章では, 聴覚心理モデル MPEG-1/Audio Psychoacoustic model 1[59, 60, 61, 62] について述べる. MPEG-1 では, SMR(signal to masking ratio) に基づき, 各周波数帯域のビット割当て数を変化させることで, 符号化を行なっている. MPEG-1/Audio Psychoacoustic model 1 は, その SMR の算出を目的とした聴覚心理モデルである. マスキングには, 周波数方向と時間方向が存在するが, この聴覚心理モデルでは, 周波数方向の滲みのみを考慮する.

SMR の算出は, 以下のステップにより行われる.

(1) 周波数変換

(2) 純音と非純音の抽出

(3) 純音と非純音, 両成分の聴覚上での音量の算出

(4) 最小可聴域と臨界帯域幅による成分の間引き

(5) マスキング閾値の計算

(6) SMR の算出

本稿では, 提案法で用いる **(1)**∼**(4)** までを説明する. また, 算出に用いる値は, サンプリング周波数が 44.1 kHz での値である.

(1) 周波数変換　まず, 入力信号をハニング窓を用いて切り出し, FFT により周波数変換する. FFT 点数は, Layer1 では 512 点, Layer2 では 1024 点. フレームシフト幅は, Layer1 では 384 点, Layer2 では 1152 点である.

(2) 純音と非純音の抽出　マスキングは, 信号の時間周波数分布によって, 時間周波数平面上での異なった広がりを持つ. この聴覚心理モデルでは, 純音と非純音の二種類を区別して, マスキング量を算出する.

純音判定は, 臨界帯域幅 (±0.5 Bark, 計 1 Bark) を考慮した上下周波数ビン (式 (A.1), (A.2)) とのエネルギの比較で行う. 比較対象に対して, 判定している成分が 7 dB 以上大きい時, 純音と判定する. また, 純音以外の成分を, 非純音とする.

純音判定に用いる帯域幅

Layer 1

$$\Delta_k = \begin{cases} \pm 2 & (2 < k < 63) \\ \pm 2, \pm 3 & (64 \le k < 127) \\ \pm 2, \cdots, \pm 6 & (127 \le k \le 250) \end{cases} \tag{A.1}$$

Layer 2

$$\Delta_k = \begin{cases} \pm 2 & (2 < k < 63) \\ \pm 2, \pm 3 & (64 \le k < 127) \\ \pm 2, \cdots, \pm 6 & (127 \le k \le 255) \\ \pm 2, \cdots, \pm 12 & (256 \le k \le 500) \end{cases} \tag{A.2}$$

(3) 純音と非純音, 両成分の聴覚上での音量の算出　純音, 非純音としての音量を算出する. 両成分は, 周波数方向への分布の広がりが異なっている. その広がりに合わせた帯域のエネルギの総和を, 聴覚上での音量とする (式 (A.3), (A.4)).

音圧の計算

純音

$$X_{tm}[k] = 10 \log_{10} \sum_{j=-1}^{1} 10^{0.1X[k+j]} \ \text{(dB)} \tag{A.3}$$

非純音

$$X_{nm}[k] = 10 \log_{10} \sum_{j} 10^{0.1X[j]} \ \text{(dB)}, \quad \forall X[j] \notin \{X_{tm}(k, k \pm 1, k \pm \Delta_k)\} \tag{A.4}$$

(4) 最小可聴域と臨界帯域幅による成分の間引き　純音, 非純音を間引きする. 間引きでは, 最小可聴域と, 臨界帯域内でのマスキングを考慮する. 各周波数ビンに対応する周波数での最小可聴域の閾値 ATH[k] は, 式 (A.5) で求められる. ここで, ATH[k] は周波数ビン k での最小可聴域の閾値である.

最小可聴域による間引き

$$\mathrm{ATH}(f) = \begin{aligned} &3.64(f/1000)^{-0.8} \\ &-6.5e^{-0.6(f/1000-3.3)^2} \\ &+10^{-3}(f/1000)^4 + 90.302 \qquad (\mathrm{dB}) \end{aligned} \tag{A.5}$$

但し, 68 dB 以上の値は 68 dB とする.
係数 90.302 は, ダイナミックレンジと最小可聴域と適合させるための補正値である.

$$P_{tm}[k] = T_{tm}[k],\ \forall k \in T_{tm}[k] \geq \mathrm{ATH}[k] \tag{A.6}$$

$$P_{nm}[k] = T_{nm}[k],\ \forall k \in T_{nm}[k] \geq \mathrm{ATH}[k] \tag{A.7}$$

続いて, 最小可聴域の閾値よりも大きい成分について, 式 (A.8), (A.9) で生成される帯域で, 式 (A.10), (A.11) のように間引きする.

スペクトルの間引き

Layer 1

$$i = \begin{cases} k & (1 \leq k \leq 48) \\ k + (k \mod 2) & (49 \leq k \leq 96) \\ k + 3 - ((k-1) \mod 4) & (97 \leq k \leq 232) \end{cases} \tag{A.8}$$

Layer 2 (ISO の Table より推測)

$$i = \begin{cases} k & (1 \leq k \leq 48) \\ k + (k \mod 2) & (49 \leq k \leq 96) \\ k + 3 - ((k-1) \mod 4) & (97 \leq k \leq 192) \\ k + 7 - ((k-1) \mod 8) & (193 \leq k \leq 464) \end{cases} \tag{A.9}$$

$$P_{tm}[i] = P_{tm}[k] \tag{A.10}$$

$$P_{nm}[i] = P_{nm}[k] \tag{A.11}$$

さらに $P_{tm}[i]$, $P_{nm}[i]$ それぞれを, 互いに重ならないように設定した臨界帯域 j 毎に総和をとり, 一つの値とする. 本稿では, この値を便宜上 $S_{tm}[j]$, $S_{nm}[j]$ とする. さらに, 信号全体での聴覚上の音量として, $S[j]$ を式 (A.12) のように定義する.

$$S[j] = 10\log_{10}\left(10^{\frac{S_{tm}[j]}{10}} + 10^{\frac{S_{nm}[j]}{10}}\right) \quad (\mathrm{dB}) \tag{A.12}$$

参考文献

[1] 松葉信彦, “ゆったりお風呂で音楽をきける防水ワイヤレススピーカー。お風呂からのプッシュトークも可能,” http://www.gizmodo.jp/2012/12/post_11417.html, 2012. (2013 年 1 月 24 日 参照).

[2] 谷井章夫, 後藤真孝, 片寄晴弘, “ミックスダウンデザインの抽出と適用,” FIT2003(情報科学技術フォーラム) 情報技術レターズ, Vol.2, pp.109–110, 2003.

[3] J. Reiss, E. Perez-Gonzalez, “Automatic Equalization of Multichannel Audio Using Cross-Adaptive Methods,” *Audio Engineering Society Convention 127*, 2009.

[4] E. Perez Gonzalez, J. D. Reiss, “Automatic Mixing: Live Downmixing Stereo Panner,” in *10th International Conference on Digital Audio Effects (DAFx-07)*, 2007.

[5] E. Perez Gonzalez, J. D. Reiss, “An automatic maximum gain normalization technique with applications to audio mixing,” in *124th AES Convention*, 2008.

[6] B. A. Kolasinski, “A framework for automatic mixing using timbral similarity measures and genetic optimization,” in *AES 124th Convention*, 2008.

[7] D. Barchiesi, J. D. Reiss, “Automatic target mixing using least-squares optimization of gains and equalization settings,” in *Proc. of the 12th Int. Conference on Digital Audio Effects (DAFx-09)*,2009.

[8] M. J. Terrell, J. D. Reiss, “Automatic monitor mixing for live musical performance,” in *Journal of the Audio Engineering Society*, 2009.

[9] E. Perez Gonzalez, J. D. Reiss, “A real-time semi-autonomous audio panning system for music mixing,” in *EURASIP Journal on Advances in Signal Processing*, 2010.

[10] D. Barchiesi, J. D. Reiss, “Reverse Engineering the Mix,” in *Journal of the Audio Engineering Society*, 2010.

[11] M. J. Terrell, J. D. Reiss, M. Sandler, “Automatic Noise Gate Settings for Drum Recordings Containing Bleed from Secondary Sources,” in *EURASIP Journal on Advances in Signal Processing*, 2010.

[12] J.D. Reiss, “Intelligent systems for mixing multichannel audio,” in *Digital Signal Processing (DSP), 2011 17th International Conference on*, July 2011.

[13] U. Zölzer, “DAFX: Digital Audio Effects,” Wiley, 2011.

[14] D. Giannoulis, M. Massberg, J. D. Reiss, “Parameter Automation in a Dynamic Range Compressor,” in *Journal of the Audio Engineering Society*, 2013.

[15] B. De Man, J. D. Reiss, “A Semantic Approach To Autonomous Mixing,” in *Journal on the Art of Record Production (JARP)*, 2013.

[16] 栗田尚之, 滝澤俊和, 行川さをり, 池田雄介, 小西雅, 山崎芳男, “多様な受聴環境を考慮したテレビ音声のミキシング - いかにして自然なデータを収集するか -,” 日本音響学会平成 14 年度秋季研究発表会講演論文集, pp.631–632, 2002.

[17] 山崎芳男, 安岡正人, 世木秀明, 宮坂栄一, 沢口真生, “高齢者に聴こえやすい放送音声サービスの研究,” 放送文化基金「研究報告」 平成 13 年度助成・援助分, 2001.

[18] 小森智康, 壇寛弥, 都木徹, 庄田清武, 黒住幸一, 小宮山摂, 星英明, 村川一広, “ラウドネスレベルを指標とした音声ミクシングバランスに関する研究,” 電子情報通信学会論文誌, Vol.J92-A, No.5, pp.344–352, 2009.

[19] N. Tsingos, “Scalable Perceptual Mixing and Filtering of Audio Signals using an Augmented Spectral Representation,” *8th International Conference on Digital Audio Effects (DAFx 2005)*, 2005.

[20] E. Gallo, G. Lemaitre, N. Tsingos, “Prioritizing signals for selective real-time audio processing,” in *Auditory Display (ICAD) 2005*, 2005.

[21] T. Moeck, N. Bonneel, N. Tsingos, G. Drettakis, I. Viaud-Delmon, D.Aloza, “Progressive Perceptual Audio Rendering of Complex Scenes,” in *Symposium on Interactive 3D graphics and games (I3D 2007)*, 2007.

[22] N. Bonneel, G. Drettakis, N. Tsingos, I. Viaud-Delmon, D. James, “Fast modal sounds with scalable frequency-domain synthesis,” in *SIGGRAPH ’ 08*, 2008.

[23] J. Mourjopoulos A. Tsilfidis, C. Papadakos. “Hierarchical Perceptual Mixing”. *Audio Engineering Society Convention 126*, 2009.

[24] P. Kleczkowski, “Selective Mixing of Sounds,” in *AES 119th Convention*, 2005.

[25] A. Kleczkowski, P. Kleczkowski, “Advanced Methods for Shaping Time-Frequency Areas for the Selective Mixing of Sounds,” in *AES 120th Convention*, 2006.

[26] P. Kleczkowski, “Selective mixing of a symphony orchestra recording,” Archives of Acoustics, Volume 31, Issue 4, 2008.

[27] P. Kleczkowski, M. Plewa, M. Pluta, “Increasing Intelligibility of Multiple Talkers by Selective Mixing,” in *AES 129th Convention*, 2010.

[28] P. Kleczkowski, “Perception of Mixture of Musical Instruments with Spectral Overlap Removed,” Archives of Acoustics, Volume 37, Issue 3, 2012.

[29] P. Kleczkowski, M. Pluta, “Perceptual Evaluation of the Effect of Threshold in Selective Mixing of Sounds,” in *Acta Physica Polonica A*, 2014.

[30] 江夏正晃, “ミックス・ダウンとマスタリングの意義,” エンジニア直伝!DAW ミックス&マスタリング・テクニック, pp.4–5, 2010.

[31] DTM・DAW の音楽制作必勝講座 ～オリジナル曲を向上させよう～, “DTM・DAW の音楽制作必勝講座 ～オリジナル曲を向上させよう～,” `http://music.sugarsword.com/`, (参照 2010-12-28).

[32] トムリンソン・ホルマン, (沢口真生, 濱崎公男, 亀川徹訳), “Sound for Film and Television 映画とテレビのための音作り,” 兼六館出版, 2000.

[33] 葛巻善郎, “エンジニアが教えるミックス・テクニック 99,” リットーミュージック, 2009.

[34] Roey Izhaki, “mixing AUDIO CONCEPTS, PRACTICES AND TOOLS.” Focal Press, 2008.

[35] 山内“Dr.”隆義, “ミックス・ダウン研究所,” Sound & Recording Magazine, Vol. 29, No. 1, pp. 154–155, 2010.

[36] 岩宮眞一郎, “図解入門よくわかる最新音響の基本と仕組み,” 秀和システム, 2007.

[37] 社団法人日本音楽スタジオ協会, “サウンドレコーディング技術概論改訂版,” 兼六館出版, 2008.

[38] “身近な計測 - 時間マスキング 2,” `http://www.onosokki.co.jp/HP-WK/nakaniwa/keisoku/masking2.htm`, (参照 2011-01-26).

[39] L McQueen, JG Terhune, “Central Masking: Fact or Artifact?,” The New School Psychology Bulletin, 2011.

[40] 難波精一郎, 苧阪直行, “音と時間,” 音響サイエンスシリーズ/ 日本音響学会編, No. 13. コロナ社, 2015.

[41] 鹿島勇, “学問の小部屋音楽部･視聴覚室～メーター編 PP メータ,” `http://kazima.pro.tok2.com/music/av/pp-meter.html` (参照 2013 年 10 月 20 日).

[42] H.H, “音量（VU）計からラウドネスメータへ,” 2011/02/25, http://www.yamaki-ec.co.jp/yamaki-onlineshop/product/link/faq_01.pdf (参照 2013年10月20日).

[43] EBU, “EBU Tech Doc 3341 ‘Loudness Metering: ‘EBU Mode’ metering to supplement loudness normalisation in accordance with EBU R 128,” 2010.

[44] 社団法人電波産業会, “デジタルテレビ放送番組におけるラウドネス運用規定,” ARIB TR-B32, 2011.

[45] 産総研, “聴覚の等感曲線の国際規格 ISO226 が全面的に改正に,” http://www.aist.go.jp/aist_j/press_release/pr2003/pr20031022/pr20031022.html (参照 2015 年 9 月 12 日).

[46] T. Matsuoka, K. Ito, “Perception of missing fundamental and consideration on its characteristics,” in *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, Vol. 3, pp. 2059–2062 Vol.3, Sept 2003.

[47] 森周司, 香田徹, 日比野浩, 任書晃, 倉智嘉久, 入野俊夫, 鵜木祐史, 鈴木陽一, 牧勝弘, 津崎実, “聴覚モデル,” 音響サイエンスシリーズ / 日本音響学会編, No. 3. コロナ社, 2011.

[48] 山内“Dr.”隆義, “ミックス・ダウン研究所,” Sound & Recording Magazine, Vol. 29, No. 3, pp. 140–141, 2010.

[49] 稲葉なるひ, “パート別リバーブ講座,” エンジニア直伝!エフェクト・テクニック基礎講座, pp. 41–56, 2004.

[50] Y. Ronen, “Vocal Clarity in the Mix: Techniques to Improve the Intelligibility of Vocals,” in *AES 139th Convention*, 2015.

[51] M. R. Schroeder, “Natural sounding artificial reverberation,” *Journal of the Audio Engineering Society*, Vol. 10, No. 3, pp. 219–223, July 1962.

[52] 角智行, “エンジニアが教えるボーカル・エフェクト・テクニック,” リットーミュージック, 2012.

[53] US 4149032 A, “Priority mixer control,” Richard W. Peters, 1979.

[54] D. Dugan, “Application of Automatic Mixing Techniques to Audio Consoles,” in AES 87th Convention, 1989.

[55] 高橋弘太, 蔦木圭悟, 吉原亨, “話速管理機能を持った原稿提示収録システム (ReCoK5) と話速バリエーション型音声データベース (SRV-DB) の公開について,” 信学技報, Vol.108, No.338, pp.227-232, 2008.

[56] 後藤真孝, 橋口博樹, 西村拓一, 岡隆一, “RWC 研究用音楽データベース : ポピュラー音楽データベースと著作権切れ音楽データベース,” 情報処理学会 音楽情報科学研究会 研究報告, 2001-MUS-42-6, 35-42, 2001.

[57] 高橋弘太, 大脇渉, “スマートミキサー — 基本原理と有効性 —,” 信学技報 , Vol.111, No. 333, pp. 43–48, 12 月 2011.

[58] 高橋弘太, 有北知弘, 大脇渉, 手塚歩, 宮地紘司, “優先度付き非線形演算による新しいサウンドミキサー — スマートミキサーの提案—,” 日本音響学会春季研究発表会講演論文集 , pp. 1035–1038, 3 月 15 日 2012.

[59] ISO/IEC, JTC1/SC29/WG11 MPEG, “Information technology—Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s-Part 3: Audio,” 1992 (”MPEG-1”).

[60] T. Painter, A. Spanias, “Perceptual Coding of Digital Audio,” Proc. IEEE, 88, pp.451-512, 2000.

[61] 山崎芳男, 金田豊, 東山三樹夫, 宇佐川毅, “音・音場のディジタル処理,” 音響テクノロジーシリーズ / 日本音響学会編 , No. 7. コロナ社 , 2002.

[62] 黒崎正行, 尾知博, “音響メディア処理と MP3,” ディジタル・デザイン・テクノロジ, Vol. 6, pp. 64–72, 2010.

[63] 古井貞熙, “新音響・音声工学,” 近代科学社, 2006.

[64] 板橋秀一編, “音声工学,” 森北出版株式会社, 2005.

[65] レイ D ケント, チャールズ リード, (荒井隆行, 菅原勉 監訳), “音声の音響分析,” 海文堂, 1996.

[66] 北村達也, 赤木正人, “単母音の話者識別に寄与するスペクトル包絡成分,” 日本音響学会誌 53 巻 3 号 , pp.185–191, 1997.

[67] ITU-T Recommendation P.800. Methods for subjective determination of transmission quality, 1996 Telecommunication standardization sector of ITU (ITU-T).

発表論文

論文

(1) 大脇渉, 高橋弘太, “時間周波数平面を用いた音声が埋もれない音信号混合法の評価,” 電子情報通信学会論文誌. (採録決定済み) (本論文の第 4 章および第 8 章に相当)

国際会議 (査読あり)

(1) W. Owaki and K. Takahashi, “Novel sound mixing method for voice and background music,” In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, pp. 290–294, April 2015. (本論文の第 4 章に相当)

研究会等

(1) 大脇渉, 高橋弘太, “スマートミキサー — 新しい音信号混合法 —,” 信学技報, Vol. 111, No. 333, pp. 37–42, 12 月 2011. (本論文の第 4.1.1.2 節に相当)

(2) 高橋弘太, 大脇渉, “スマートミキサー — 基本原理と有効性 —,” 信学技報, Vol. 111, No. 333, pp. 4348, 12 月 2011.

(3) 大脇渉, 有北知弘, 手塚歩, 宮地紘司, 高橋弘太, “音声を埋もれさせない音信号混合法— スマートミキサーの提案 —,” 日本音響学会春季研究発表会 講演論文集, pp. 699–702, 3 月 2012. (本論文の第 4.1.1.3 節に相当)

(4) 高橋弘太, 有北知弘, 大脇渉, 手塚歩, 宮地紘司, “優先度付き非線形演算による新しいサウンドミキサー— スマートミキサーの提案 —,” 日本音響学会春季研究発表会講演論文集, pp. 1035–1038, 3 月 2012.

(5) 大脇渉, 有北知弘, 宮地紘司, 高橋弘太, “スマートミキサー— 聴覚心理モデルに基づいたゲインマスク生成 —,” 信学技報, Vol. 112, No. 347, pp. 47–52, 12 月 2012. (本論文の第 4.2.1 節に相当)

(6) 大脇渉, 有北知弘, 宮地紘司, 高橋弘太, “スマートミキサー — 時間周波数平面上での位相操作 —,” 信学技報, Vol. 112, No. 478, pp. 13–18, 3 月 2013. (本論文の第 4.2.2, 4.2.3 節に相当)

(7) 宮地紘司, 大脇渉, 高橋弘太, “音響圧縮形式の時間周波数解析を利用した音信号混合法,” 信学技報, Vol. 113, No. 503, pp. 49–53, 3 月 2014.

(8) 有北知弘, 大脇渉, 宮地紘司, 高橋弘太, “急峻な周波数変化に対応した信号解析法とスマートミキサーへの応用,” 信学技報, Vol. 113, No. 503, pp. 55–60, 3 月 2014.

(9) 田本祐樹, 大脇渉, 高橋弘太, “画像を媒介とした直観的操作が可能なスマートミキサー,” 信学技報, Vol. 114, No. 274, pp. 15–19, 10 月 2014.

(10) 旭岡舜介, 大脇渉, 高橋弘太, “フォルマント構造維持を規範とした音声信号混合法,” 信学技報, Vol. 114, No. 274, pp. 21–26, 10 月 2014.

(11) 大脇渉, 旭岡舜介, 高橋弘太, “フォルマント構造維持を規範とした音声信号混合法, ” 情報処理学会 音楽情報科学研究会 研究報告, vol. 2015-MUS-107, No. 66, pp. 1–4, 5 月 2015. (本論文の第 5 章に相当)

(12) 大脇渉, 高橋弘太, “歯擦音と母音の識別性を重視する音声信号混合法, ” 信学技報, vol. 115, No. 126, pp. 7–10, 7 月 2015 年. (本論文の第 6 章に相当)

(13) 長谷川政良, 大脇渉, 高橋弘太, “騒音下における音声再生の識別性を重視したスマートミキサー,” 日本音響学会 騒音・振動研究会資料, Vol. 102, No. 10, 8 月 2015.

(14) 池田友和, 大脇渉, 高橋弘太, “楽音のスペクトル構造に基づいた音信号混合法, ” 情報処理学会 音楽情報科学研究会 研究報告, vol. 2015-MUS-109, No. 2, pp. 1–6, 11 月 2015.

特許

(1) 特許第 5057535 号, “ミキシング装置、ミキシング信号処理装置、ミキシングプログラム及びミキシング方法”, 高橋弘太, 大脇渉 (2011).

(2) US20140219478 A1, “MIXING DEVICE, MIXING SIGNAL PROCESSING DEVICE, MIXING PROGRAM AND MIXING METHOD”,
Takahashi, Kota, Owaki, Wataru (2012). (PCT 出願済み, 米国への移行手続き中)